Gene regulatory networks during zebrafish endoderm development

Mariossi, Andrea

Awarding institution:
King's College London

Gene regulatory networks during zebrafish endoderm development

Andrea Mariossi

A Dissertation submitted for the

Degree of Doctor of Philosophy
to

Randall Centre for Cell & Molecular Biophysics

Faculty of Life Sciences and Medicine

March 2019

Abstract

During development, gene regulatory networks control cell fate decisions and differentiation leading to formation of the adult organism. A critical part of this process is the establishment of the three germ layers – ectoderm, mesoderm and endoderm. The endoderm, the inner germ layer, gives rise to the respiratory and gastrointestinal tracts and contributes to organs such as the pancreas, liver and lungs. The establishment of this germ layer is a process dependent on the integration of multiple transcriptional and signalling inputs, and while the formation of ectoderm and mesoderm have been studied extensively, our knowledge about the regulatory network controlling endoderm development is still limited. The aim of my PhD is to investigate the network of transcription factors that determine the specification of endodermal lineage in zebrafish and how these factors integrate and interact to bring about correct gene expression. In order to do so, I combine both experimental and computational methodologies. Firstly, I focus on the role of Sox and Mix family members and their interactions with binding partners in specifying and patterning endodermal cells during gastrulation. I use ChIP-exo, a form of chromatin immunoprecipitation combined with exonuclease digestion and high-throughput sequencing, to try and identify genomic positions bound by the different transcription factors. I then isolate endodermal cells by flow cytometry using a transgenic line carrying a fluorescent reporter gene, *tg*(*sox17:GFP*), in order to characterise the transcriptomic signature of developing endodermal cells. Finally, I combine the data generated together with existing published data sets to construct the gene regulatory network underpinning endodermal fate in zebrafish. Together, these findings extend our understanding of embryonic endodermal development, and move us one step closer to reconstructing the pattern of genetic regulation that happens during the developmental specification of endoderm. This knowledge will ultimately help us to formulate a comprehensive map detailing how a pluripotent cell in the zebrafish embryo becomes committed to an endodermal fate.

Acknowledgments

This work would not have been possible without the help and support of many people. Firstly, I would like to thank my supervisor, Dr Fiona Wardle for giving me the opportunity to do my PhD research in her lab and for supervising me throughout. Thank you very much to members of the Wardle lab past and present, particularly Steve for teaching and guiding me at the start of my PhD and Husam, who introduced me to bioinformatics. Sarah and Reuben kept me sane during long hours in the lab for which I am forever grateful. Thanks also to my second supervisor, Professor Anthony Graham and members of my thesis committee, Dr Baljinder Mankoo, Dr David Fear and Professor Sasi Conte for their continued support throughout my PhD and their invaluable comments and feedback.

My PhD was part of a Marie Curie ITN and as such, I was lucky enough to travel around Europe and meet other highly motivated PhD students who became both my advisors and my friends. In particular, thanks to Ana and Claudia for the time spent in Liege and for the introduction to FACS. Huge thanks also go to Matthias and Kadir at the Karolinska Institutet for being so patient with me, answering my questions regarding the data analysis and helping me learn bioinformatics.

I had the privilege of sharing office space at King's College with members of the Zammit, Hughes and Hinits labs; more specifically Tapan, Chris, Pedro, Maryna, Mike, Fips and Ailbhe. We had lots of fun together over the years and I will always look back fondly on the time we spent together. Two outstanding post-docs, Dr Nicolas Figeac and Dr Massimo Ganassi, were instrumental in supporting me throughout my project, both academically and personally and I am so grateful to them.

Thanks also to members of the Smith lab at the Crick, particularly Kayleigh, Fay, Clara, Camille, Kenzo, George and Luis for putting up with me whilst writing up and for helpful comments and advice.

Without the help of an astonishingly bright scientist in my life, my best friend Johanna Prueller, none of this would have been possible. You know how much I value your opinion because you know everything. And you actually do. Thank you so much.

To the Italian gang: Andrea and Luca, my skiing and scuba diving friends, you have been amazing, presenting me with multiple opportunities to have fun and relax throughout the journey of my PhD. A big thank you goes to the old crew, Andrea, Laura, Gregorio, Viviana and John for the enthusiastic support in both small and big ways.

Lastly but certainly not least, thank you to my family, mum and Grandma. You have been there from the beginning with unwavering support and patience and I promise this is my last degree. To my girlfriend and sweetheart Rebecca. This PhD has been a journey and you have been there throughout the ups and downs. You made me carry on when I was down, acknowledge my strengths when I was on top and above all, you believed in me. In turn, I grew to believe in myself and learned to never give up. You inspire me and I love you. Here's to our next chapter.

Declaration

The work presented in this thesis is entirely my own work, apart from that described below:

- Phylogenetic trees and radioactive immunoprecipitation assays were carried out by Amanda Evans.

The thesis has not been accepted in any previous applications for a degree and all sources of information have been acknowledged.

Andrea Mariossi

Table of contents

List of figures

List of tables

Abbreviation

Bmp:bone morphogenetic protein

Bp:biological processes

Cc:cellular components

ChIP-exo: chromatin immunoprecipitation combined with lambda exonuclease

ChIP-seq: chromatin immunoprecipitation follow by sequencing

CRISPR/Cas9: clustered regularly interspaced short palindromic repeats/crispr-associated protein 9

DEGs: differentially expressed genes

DFCs: forerunner cells

EGFP: enhanced green fluorescent protein

FACS: fluorescence-activated cell sorting

FDR:false discovery rate

Fgf:fibroblast growth factors

Fox:forkhead box

GFP: green fluorescent protein

GO: gene ontology

GRN: gene regulatory network

HEK293: human embryonic kidney cell line

Hh:hedgehog signalling

Hi-C: chromosome conformation capture

HMGB :high-mobility group

Hpf: hour post fertilization

HRM: high resolution melting

IDR: irreproducibility discovery rate

IHC: immunohistochemistry

IP: immunoprecipitation

KEGG:KYOTO encyclopaedia of genes and genomes

Kv: kupffer's vesicle

Log FC: log fold change

MBT: midblastula transition

MF: molecular function

MFI: mean fluorescence intensity

MNase: micrococcal nuclease

MO: morpholino

Mxtx: mix-type homeobox gene

NGS: next-generation sequencing

PE: paired-end

Pou: pit-oct-unc

PWM: Position weight matrix

RA: retinoic acid

RPM: reads per million

RRL: rabbit reticulocyte lysate translation systems

scRNA-seq: single cell rna-sequencing

SE: single-end

Sry:sex-determining region y

TADs: topologically associated domains

TF: transcription factor

TPM: transcripts parts per million

WB: western blotting

WISH: whole mount in situ hybridisation

FZIN: Zebrafish Model Information Network

ZGA: zygotic genome activation

λ-exo: lambda exonuclease TPM: transcripts parts per million

Preface

My work since the beginning of my PhD in October 2015 has focused on the molecular program controlling the formation of endoderm using zebrafish as a model system and combining a mix of experimental, genomic and computational techniques. My thesis consists of four main results chapters that address my approaches, the underpinning rationale and my findings.

The first result chapter (Chapter 3) focuses on two Sox genes known to be involved in endoderm formation, s*ox17* and *sox32* and their respective genetic regulatory functions, with particular emphasis on *sox32*. In addition, the roles of other known endodermal genes such as Mix paired-like homeobox genes 1 (*mixl1*) are considered from a genome wide perspective, to identify both novel binding partners and components of the endodermal gene regulatory cascades that involve Sox32 and Mixl1 proteins.

Chapters four and five describe my interest in transcriptomics and the way in which it has revolutionised our understanding of how genomes are expressed. Here, I compare transcriptomic data from wild type and mutant zebrafish (*sox32*$^{-/-}$ and *mixl1*$^{-/-}$) to come to understand how the loss of specific endodermal genes during gastrulation perturbs the formation of endoderm derived organs. Additionally, I also detail the characterisation of the previously described *sox17:GFP* line that I used during these experiments, as it demonstrated genetic inconsistencies. As well as being used for this project, the *sox17:GFP* line is frequently used; I therefore considered it both pertinent and important to the field of zebrafish development to describe the anomalies I identified with this line hosted at King's College fish facility.

The last chapter details my interest in combining next generation sequencing data with data obtained from more classical genetic approaches in order to create an endodermal gene regulatory network (GRN). In particular, I combine lessons learned from studies in early vertebrate embryos such frogs and mice, and from studies in human embryonic stem cells to identify critical network interactions in the developing zebrafish embryo. This approach has provided me with a deeper insight into the mechanisms of endodermal gene regulation in the early zebrafish embryo and ultimately the GRN governing endoderm differentiation.

Chapter 1 – Introduction

Embryonic development has always been one of the great sources of wonder for biologists. How can a single fertilised cell go on to develop into an entire organism composed of different specialised cells organised into multiple tissues and organs? Development of an organism is a generative program, whereby the instructions encoded within the DNA are read, interpreted and implemented in order to build the organism. The key questions are: How are all the commands for the formation of the adult organism progressively interpreted in any particular cell? How are the lineages at the different developmental stages determined, and how do the cells respond to other neighbouring cells and their wider environment? Over the past few decades, the study of developmental biology has uncovered some of the secrets as to how this process works and amongst the other major milestones achieved, has now started to describe the progressive changes that cells undergo during development, including the complex interactions of different signalling pathways and transcription factors that together, form the regulatory networks underpinning cell fate decisions (Davidson et al., 2002; Peter, 2017; Cholley et al., 2018). This is providing us with an ever-clearer picture of the processes behind development itself, and furthermore, aging, healing, and disease.

All vertebrates are characterised by the same fundamental organisation of cells in the early embryo, whereby pluripotent and multipotent progenitors are transformed into specific cell types in spatially stereotypic arrangements (Salazar-Ciudad et al., 2003; Schier and Talbot, 2005; Briggs et al., 2018). These regulatory programs lead to the division of the embryo into three germ layers: the outer layer called ectoderm, the middle layer named mesoderm, and the endoderm, which is the inner most layer of the embryo. The ectoderm gives rise to the epidermis and nervous system. The mesoderm gives rise to muscle, blood, dermis, bones, gonads, kidneys and connective tissues. The endoderm is associated with the most internal organs, creating the epithelium of the digestive tract and respiratory system and organs associated with the digestive system, such as the liver and the pancreas. The organisation of the embryo into three germ layers occurs during a process called gastrulation, from the Greek word gut (gastḗr). During gastrulation, cell movements reshape and reposition the cells in order to segregate into the three germ layers (Warga and Kimmel, 1990; Gritsman et al., 2000; Keller, 2005).

The cell movements that occur during gastrulation and the segregation of cells into the three different germ layers (and indeed the cells' future identities), are governed by the underpinning gene regulatory networks (GRNs). GRNs are essentially maps that detail how cells diversify their responses to the inputs and outputs that occur during different scenarios, dependent upon the components regulating a developmental program. A GRN details the interactions between genes, transcription factors and other components of signalling pathways that form an intricate network-like architecture that defines the identity and function of a cell. GRNs exemplify the changes in cellular competence (the ability to respond) to a given signal over time and they thus provide the basis on which to model the dynamics of signalling. GRNs demonstrate how transcriptional networks are employed to convert signals into stable patterns of gene expression and they can be used to predict how the interactions between the different components evolve along the developmental stages. GRNs therefore provide an explanation of the genetic interactions that drive differentiation and development (Levine and Davidson, 2005; Li and Davidson, 2009).

Much is known about the GRNs underpinning ectodermal and mesodermal fate, yet the processes leading to endodermal cell identity are more poorly understood. Despite this, studies have begun to elucidate an evolutionarily conserved molecular pathway that specifies the endoderm during gastrulation, and this information has also helped in our ability to differentiate human endoderm tissue from stem cells (Mohammadnia et al., 2016; Yiangou et al., 2018; Chia et al., 2019), a finding that may have broad implications for disease and bridging the gap between *in vitro* and *in vivo* models.

In this introduction to my PhD thesis, I first describe the importance of fate choice decision in the early steps of embryo development and explain the importance of GRNs. I then review the early steps of zebrafish embryo development, from fertilisation to gastrulation. I then describe what is known from the existing literature about the network of transcription factors (TFs) that determines the specification of the endodermal lineage, and how these factors integrate and interact during gastrulation. Understanding the development of endoderm and dissecting the GRNs that ultimately create endoderm-associated organs could provide us with important insights into diseases associated with these systems and tissues (Ober and Grapin-Botton, 2015). Endoderm formation is a critical embryonic event and its study has and will continue to provide thoughtful insight into fundamental developmental mechanisms that could

aid in developing successful molecular therapeutics for human benefit (Ober and Grapin-Botton, 2015; Yiangou et al., 2018).

1.1 Waddington's landscape and the different shades of development

In 1940, Conrad H. Waddington introduced the famous 'epigenetic landscape' metaphor, a new conceptual view representing embryonic development in which he described the various paths a cell might take during cell differentiation and organism development (Waddington, 1940) (Figure 1.1) The landscape illustrates how a cell during embryonic development negotiates a cascade of branching lineage choices, avoiding alternative fates at each juncture to conclusively commit to a single lineage.

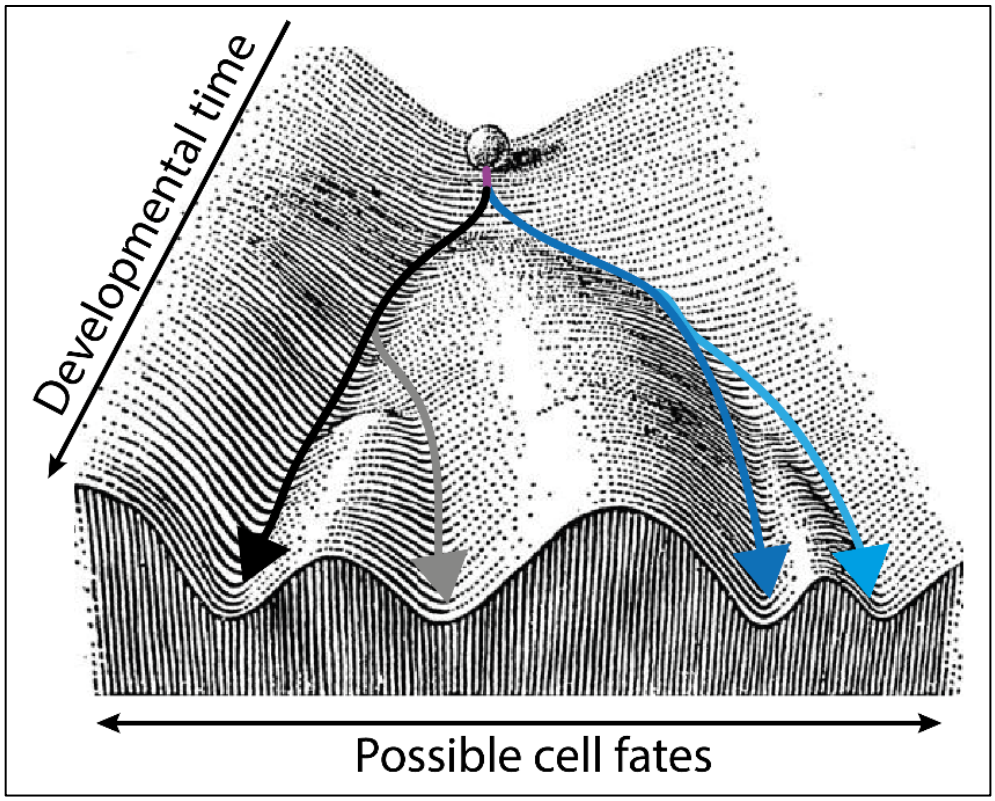


Figure 1.1 Waddington's landscape provides a simple, visual representation of the concept of cell differentiation during development. The low points (valleys) of the landscape signify developmental pathways that a cell can travel along and the high points define the limits of the pathways, symbolising regulation. All pathways branch from a single perpendicular line (purple; a toti-, pluri- or multipotent cell) and represent the different trajectories a cell may take during differentiation (black, grey, dark and light blue arrows). The features of the landscape, including the hills and valleys and the points at which each path branches, are important to Waddington's argument. At the outset of development, new valleys represent alternative cell fates, and ridges keep cells from switching fates. Figure shows the original artwork of Waddington (1940) adapted by the addition of arrows and labels.

In his drawings, a cell, represented by a ball, sits at the top of a hill. As it rolls down the hill the ball faces a landscape of uneven slopes and bifurcating valleys that ultimately direct the ball into different possible paths on its way to a final destination. In this metaphor, the hill represents the cell differentiation process and each new valley along the path represents a cell fate decision the cell can make. The ridges between the valleys maintain the cell trajectory once it has been chosen. All these hills and valleys ultimately channel the ball towards a specific position at the bottom of the hill. This simple rendering of the epigenetic landscape as a hill means that the top represents the toti/pluri/multipotent state of a cell and the positions at the bottom represent the different types of fully differentiated cells.

Beyond its misleading simplicity, Waddington included concepts such as regulation, competence and induction in his model. In support of this idea, multiple studies in the last eight decades have expanded on this model and incorporated new information into the system he proposed. This further refinement of Waddington's branched track diagram currently incorporates the following: (i) the gradual decrease in potency during development that Waddington illustrated by the tilt of the landscape, is now understood to be caused by the loss of valleys as opposed to the creation of new ones; pluripotent cells have all the landscape available to them and along the canalisation to a fate, the valley that is not chosen disappears; (ii) the epigenetic barriers between sharply distinct cell fates, depicted as the hills between the valleys, can be overridden (Ferrell, 2012; Baedke, 2013). Despite the successes in charting lineage intermediates in tissues, key lineage branchpoints remain controversial and it has so far not been possible to systematically identify the regulatory mechanisms that control cell fate at each branchpoint.

1.2 Gene regulatory networks

Embryonic development is established by the coordinated and precise regulation of thousands of genes that drive appropriate developmental fates through GRNs. There are different layers of players that are able to influence gene regulation within these networks, from TFs that bind specifically to regulatory regions (promoters and enhancers) and promote transcription, to regulators of chromatin structure and accessibility such as histone modification enzymes. Each GRN contains the logic circuits that describe the interactions between the cis-regulatory regions of genes and TFs for a developmental process at a given time and in a specific space. GRNs encode the response of a cell to signalling pathways and other

developmental signals, instructing the cell to adopt a particular developmental trajectory and ultimately become part of a defined tissue or organ.

If for one moment we return to the concept of Waddington's landscape, the previous figure (Figure 1.1) is complemented by a second image exposing the underside of the landscape. As depicted in Figure 1.2, the landscape is tied by strings to a matrix of pegs. The strings represent the GRN in which the genes interact with each other in complex relationships to shape the surface of the landscape and therefore dictate the developmental path of a cell.

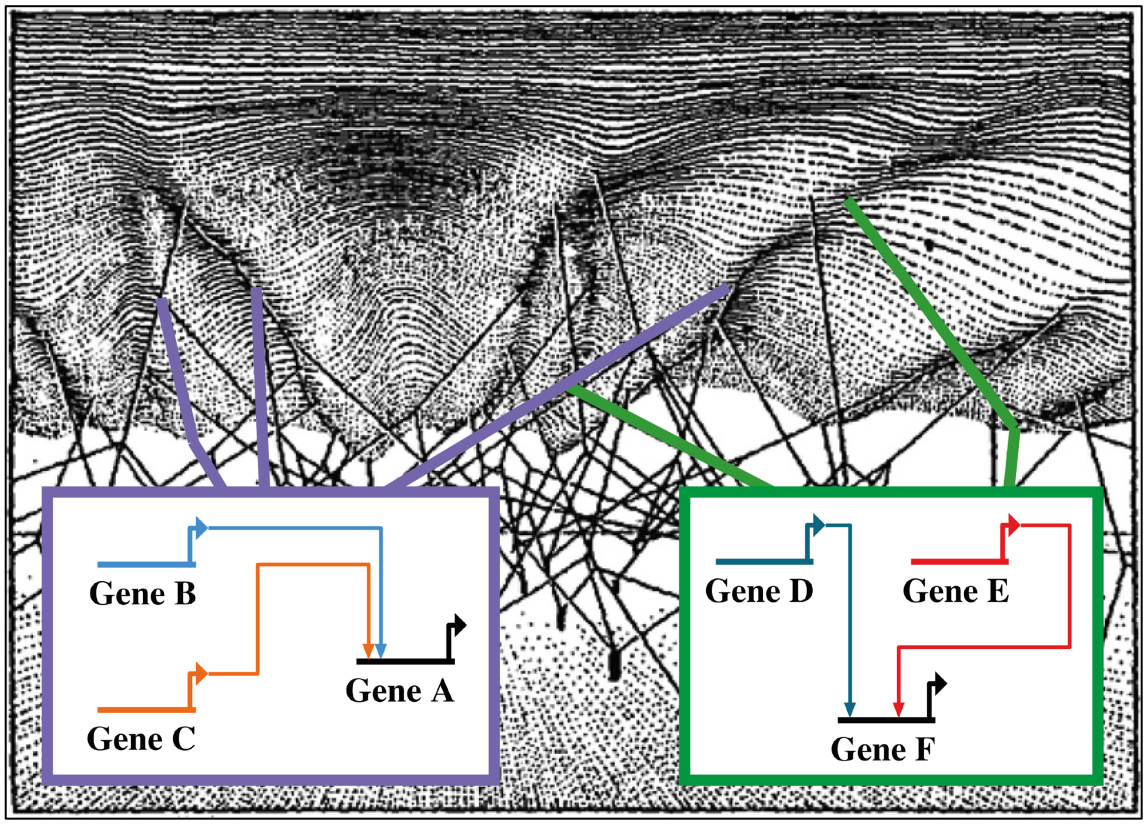


Figure 1.2 Second part of Waddington's landscape. Embryonic development is a progressive program in which a single totipotent cell gives rise to hundreds of distinct differentiated cell types, each of which has a specific regulatory network that control it (small boxes at the bottom). The topography of the landscape arises as a result of the interactions between the genes and the surface. The genetic interactions are the major determinants of the landscape's shape (coloured strings).

Traditionally, the analysis and reconstruction of a GRN has been done using perturbation experiments, where changes that occur due to disruption of a particular component of the network are mapped (Li and Davidson, 2009). This perturbation may lead to a small or wide range of adjustments in the network itself. After introducing a change in the system, either *a priori* knowledge or a speculative hypothesis regarding what kind of effects the perturbated

gene may cause is needed to test the change. However powerful and precise this approach is, it cannot capture the entire biological complexity of most systems, nor can it integrate the information across the multiple regulatory layers that the perturbation has achieved. More recently therefore, 'omics' technologies have emerged as a system-wide approach to provide a more comprehensive view of the interaction within a GRN.

In the last decade, high-throughput sequencing technology, also known as next-generation sequencing (NGS), has revolutionised the study of genomics, epigenetics and transcriptomics, providing powerful insights in the molecular landscape and associated biological processes. The advantages of NGS are the speed, cost-effectiveness, the huge amount of data that can be obtained and the ability to survey global expression patterns quickly (Buermans and den Dunnen, 2014; Ari and Arikan, 2016). Many previously existing techniques (both for RNA and DNA) have been coupled to high-throughput genome wide sequencing, resulting in pioneering studies into a variety of biological questions in humans and mouse, ranging from the sequencing of whole genomes and chromatin analyses, to the discovery of new TF binding sites and RNA expression profiling. Consequently, several zebrafish laboratories have started adopting an 'omics approach' in order to have a comprehensive, or global, assessment of the entire system and then focussed on the study of gene regulation and comparative evolutionary genomics. An exponential number of studies have been published, including transcriptomic analysis (mRNA and long noncoding RNA-seq) (Aanes et al., 2011; Vesterlund et al., 2011; Pauli et al., 2012; Harvey et al., 2013; Yang et al., 2013; Junker et al., 2014; White et al., 2017), epigenetic analysis (histone ChIP-seq) (Aday et al., 2011; Lindeman et al., 2011; Bogdanovic et al., 2013), ChIP-seq for sequence-specific TFs (Xu et al., 2012; Leichsenring et al., 2013; Winata et al., 2013; Nelson et al., 2014; Nelson et al., 2017; Lukoseviciute et al., 2018), ribosomal profiling (Bazzini et al., 2012), measurement of chromatin accessibility (ATAC-seq) (Fernandez-Minan et al., 2016; Kaaij et al., 2016; Quillien et al., 2017), nucleosome organisation and three-dimensional architecture of genomes (Hi-C) (Kaaij et al., 2018). These studies have provided a very useful framework for the identification of novel noncoding DNA elements, regulatory sequence features and TFs that drive gene expression dynamics during zebrafish development. In particular, they have proved very valuable to our understanding of the effects of nonlinear interactions, such as those produced by the combinatorial actions of TFs, as well as to predict the genetic responses in normal and mutant organisms.

GRNs have been successfully used to describe biological processes in several organisms and systems, including the yeast *Saccharomyces cerevisiae* (Lee et al., 2002), the mammalian immune system (Singh et al., 2014), the plant *Arabidopsis thaliana* (Azpeitia et al., 2013), the sea urchin (Levine and Davidson, 2005; Peter and Davidson, 2010), the sea squirt *Ciona* (Imai et al., 2009) and the fruit fly *Drosophila* (Levine and Davidson, 2005). In zebrafish, GRNs have been used to describe the neural crest (Williams et al., 2018), melanocyte pigmentation (Greenhill et al., 2011), iridophore cells fate (Petratou et al., 2018) and mesendoderm specification (Morley et al., 2009; Nelson et al., 2017). However, a GRN has not yet been described that bridges the gap between the mesendodermal population and the specified endodermal cells.

1.3 Early stages of zebrafish embryo development

During vertebrate development, endodermal cells form a broad range of tissues and organs, including the digestive system and associated organs such as the liver, pancreas, thymus and thyroid as well as the epithelial lining of the respiratory tract (lungs for mammals and gills for fish) (Stainier, 2002; Zorn and Wells, 2009). Endoderm specification starts during gastrulation, and the key factors involved in this process are regulated by the Nodal family of signaling molecules (Feldman et al., 1998). Various TFs and signaling molecules regulating endoderm development in vertebrates have been identified, broadly delineating the developmental landscape (Schier et al., 1997; Alexander et al., 1999; Reiter et al., 1999; Warga and Nusslein-Volhard, 1999; Feldman et al., 2000; Kimelman and Griffin, 2000; Poulain and Lepage, 2002; Dougan et al., 2003; Reim et al., 2004; Bjornson et al., 2005; Poulain et al., 2006). Yet, the conserved signaling and transcriptional pathways responsible for endoderm formation are not understood in enough depth to allow full recapitulation of the cascade of events that leads to the definition of endoderm.

Zebrafish embryo development can be divided into 7 stages – zygote, cleavage, blastula, gastrula, segmentation, pharyngula, and hatching period (Kimmel et al., 1995). After fertilisation, the zygote undergoes multiple cell divisions or cleavages from 0 - 2.25 hours post fertilisation (hpf); during the blastula period (from 2.25 hpf - 5.25 hpf), continuous divisions produce a cap of cells called the blastodisc, which sits on top of the yolk mass. At this stage, cells are not determined to a specific developmental fate, they can all contribute to all tissues of the forming embryo (Zorn and Wells, 2009). During the gastrula stage from 5.25 to 10 hpf,

the cell fates becomes restricted at the tissue level, which results in the formation of the three aforementioned germ layers: ectoderm, mesoderm, and endoderm. By the end of gastrulation, cells have migrated and are positioned in the respective territories and cells become committed to specific fates (Ho and Kimmel, 1993). Thereafter, between 10 and 24 hpf, segmentation and somitogenesis are both completed.

At the midblastula stage (~4.3 hpf), between the border of the yolk cell and the animal pole of the embryo, a yolk syncytial layer (YSL) is formed, and epiboly begins. The YSL is a single row of cell nuclei along the blastodisc margin, created by fusion of the collapsing marginal blastomeres with the yolk mass. The YSL adopts a crucial function in the induction of the germ layers, when it secretes the primary signal for mesendoderm specification (Nodal and obligatory cofactors) and supply the maternal and zygotic TFs (Mxtx2, Nanog, Pou5f1). In addition, there is a single cell layer that covers the surface of the blastoderm, referred to as the layer of enveloping cells (EVL) (Figure 1.3).

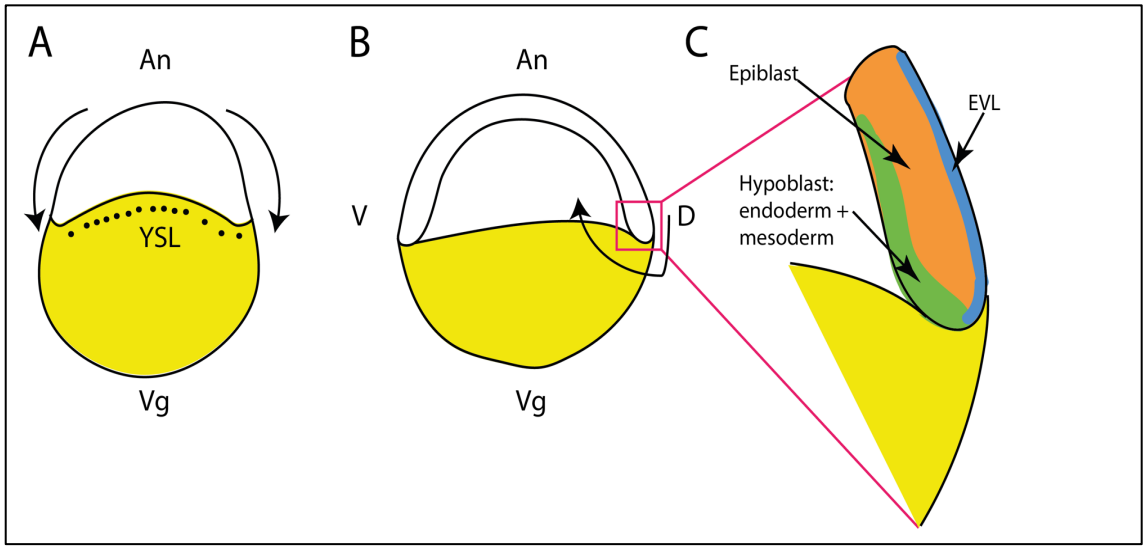


Figure 1.3 Zebrafish embryos at the onset of epiboly. (**A**) At 4.3 hpf, the blastoderm begins to move around the yolk cell in a process called epiboly. Cells at the blastoderm margin fall into the yolk cell and form the yolk syncytial layer (YSL). The extraembryonic enveloping layer (EVL, blue) covers the blastoderm. Black arrows refer to cell movements. (**B**) At 5.25 hpf, at the dorsal side of the embryo, signals from YSL induce invagination of the blastoderm. The ingression of the cells at the margin which will give rise to mesodermal and endodermal structures later on form the embryonic shield. (**C**) Zoomed diagram of the situation at 5.25 hpf. Epiblast tissue is orange, mesoderm and endoderm in green, EVL in blue and yolk in yellow. An: animal, Vg: vegetal, V: ventral, D: dorsal. Modified from Solnica-Krezel (2002).

Epiboly and gastrulation are the processes during which the main axes of the embryo are created; anterior-posterior and dorso-ventral. During epiboly (which starts at ~4 hpf) the YSL and the blastoderm spread over the yolk cell (Figure 1.3A, arrows). Epiboly is described in percentages, as it indicates by how much the spreading cells have surrounded the yolk mass. When the leading edge of the blastoderm reaches the equator of the embryo, at 50% epiboly (5.25 hpf), gastrulation begins and cells involute around the margin (the interface of the blastoderm and yolk cell) to form the hypoblast, while the outer cells form the epiblast. As a consequence of this involution of cells, a local thickening known as the embryonic shield, appears at the margin of the embryo. The cells at the blastoderm margin which involute first will give rise to mesodermal and endodermal derivatives, cells that are more distant from the margin and thus involute later form only mesoderm, while non-involuting cells farthest from the margin form ectoderm (Figure 1.3C). Epiboly is complete at the end of the gastrula period when the YSL, EVL and blastoderm have engulfed the entire yolk mass.

1.4 Endoderm spatial domain

Endoderm development begins with Nodal signal diffusion from the YSL in the ventral and lateral margin of the blastula at around 4.33 hpf (midblastula) and the induction of a transitory cell population, denominated mesendoderm, which, as the name suggests, has the potential to differentiate into both mesoderm and endoderm cell lineages (Figure 1.4) (Erter et al., 1998; Dougan et al., 2003; Schier, 2003; van Boxtel et al., 2015). Future endodermal cells are the first to involute during epiboly and they individually migrate anteriorly to create a monolayer, which then consolidates along the dorso-ventral axis in a topographic arrangement of the future digestive system (Warga and Nusslein-Volhard, 1999). Mesoderm arises from the cells that have ingressed between the endoderm and the non-involuting epiblast (Kimmel et al., 1990; Warga and Kimmel, 1990). Cell lineage tracing studies have shown that the most ventral endodermal cells generate the alimentary canal, whereas the most dorsal cells yield the anterior endoderm, the pharynx/gills, whilst the lateral cells will give rise to the digestive organs, including the liver and pancreas, the latter of which generates exocrine and endocrine constituents (Warga and Nusslein-Volhard, 1999).

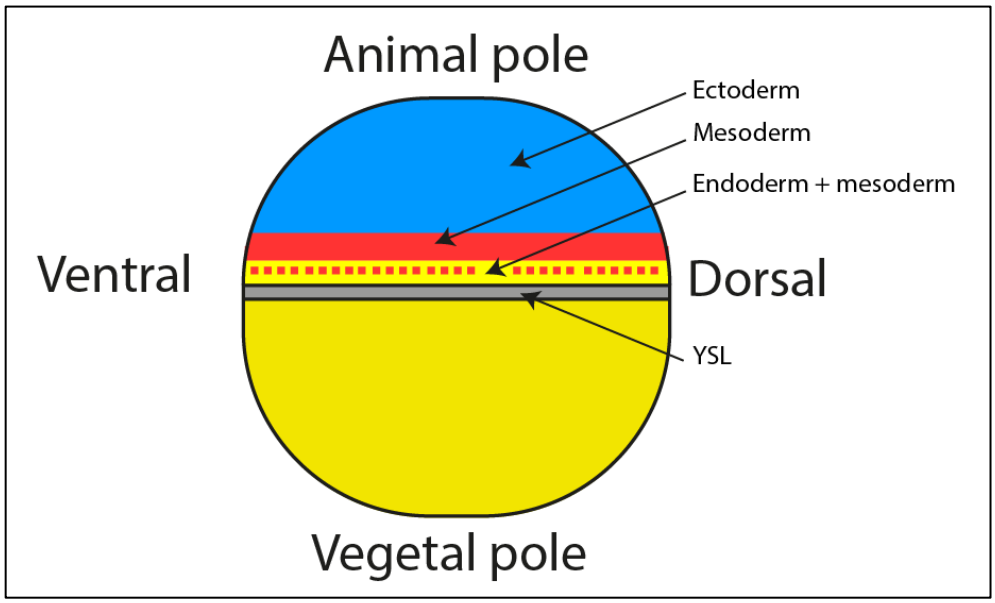


Figure 1.4 Cell lineages projected on to a schematic of a zebrafish embryo at blastula stages. At the onset of epiboly, the mesendoderm territory is situated above the YSL. At the midblastula stage, the animal pole cells will give rise to the ectoderm (blue), whereas marginal cells will form both mesoderm and endoderm (red); these two lineages have not yet separated. This separation occurs during the late blastula period.

1.5 Nodal signalling initiates endoderm and mesoderm formation

Our understanding of vertebrate endoderm formation is derived from studies primarily in *Xenopus*, mouse and zebrafish. These studies have highlighted the essential and conserved role of Nodal signalling in all vertebrates for appropriate induction and development of endoderm and mesoderm in the right place at the right time (Varlet et al., 1997; Feldman et al., 1998; Osada and Wright, 1999; Schier and Shen, 2000), gastrulation movements (Feldman et al., 2000; Pézeron et al., 2008; Liu et al., 2018), and control of left-right axis patterning (Schier, 2003; Schier and Talbot, 2005).

Nodals are members of the TGF-β superfamily of signalling factors which are highly conserved in vertebrates. They are required for the formation of the anterior-posterior axis of the embryo and for the specification of both mesoderm and endoderm as mentioned above. Starting from blastula stage Nodal ligands are expressed at the margin and YSL creating a morphogen gradient along the vegetal-animal axis (Chen and Schier, 2002; Fan et al., 2007). Nodals bind to type I and II TGF-β receptors on the surface of cells followed by phosphorylation of intracellular signal transducers, Smads 2 and 3 (Schier and Shen, 2000; Schier, 2003). The type I TGF-β receptor in zebrafish is Acvr1ba (TARAM-A/Tar) (Aoki et al., 2002b).

In addition to type I and type II receptors, Nodal signalling requires EGF-CFC coreceptors. Coreceptors of the EGF-CFC family are known as TDGF1 and CFC1 in mouse (previously known as Cripto and Cryptic respectively) and are essential for Nodal signalling (Ding et al., 1998). In zebrafish, type II TGF-β receptor activation induces interaction with the EGF-CFC coreceptor, called Tdgf1 (previously known as One-eyed-pinhead) (Gritsman et al., 1999). Nodal signalling is antagonised by feedback inhibitors, such as Lefty proteins, which are themselves members of the TGF-β family, which block EGF-CFC coreceptors and enhance the degradation of type I receptors (Chen and Schier, 2002) (Figure 1.5).

Activation of Nodal receptors leads to the phosphorylation of Smad2/3 which then bind to Smad4 forming a heterodimer, followed by the translocation of the complex to the nucleus (Dick et al., 2000; Weng and Stemple, 2003) where it interacts with TFs to activate target gene expression. In zebrafish, key downstream targets are *foxh1*, *mixl1* and *sox32* with the latter playing an essential cell autonomous role in endoderm formation (Kikuchi et al., 2001).

As mentioned above, Nodal signalling leads to phosphorylation of Smad2/3. *Smad2* null mice have a strong embryonic phenotype whereas *Smad3* knockout mice develop normally. This can be explained by different temporal regulation of the two SMADS, with SMAD2 being more abundant than SMAD3 in the early stages (Robertson, 2014). A similar pattern was observed in *Xenopus* (Howell et al., 2001) and zebrafish where levels of Smad3 are low in early development. Recently, Dubrulle et al. (2015) and Nelson et al. (2014) reiterated the importance of Smad2 for Nodal signalling and mesendoderm specification in zebrafish; the maternal-zygotic *smad2* mutant (MZ*smad2*) phenocopies a loss of Nodal signalling, showing that in zebrafish, as in the other two model species, Smad3 plays a minimal role at the midblastula stage.

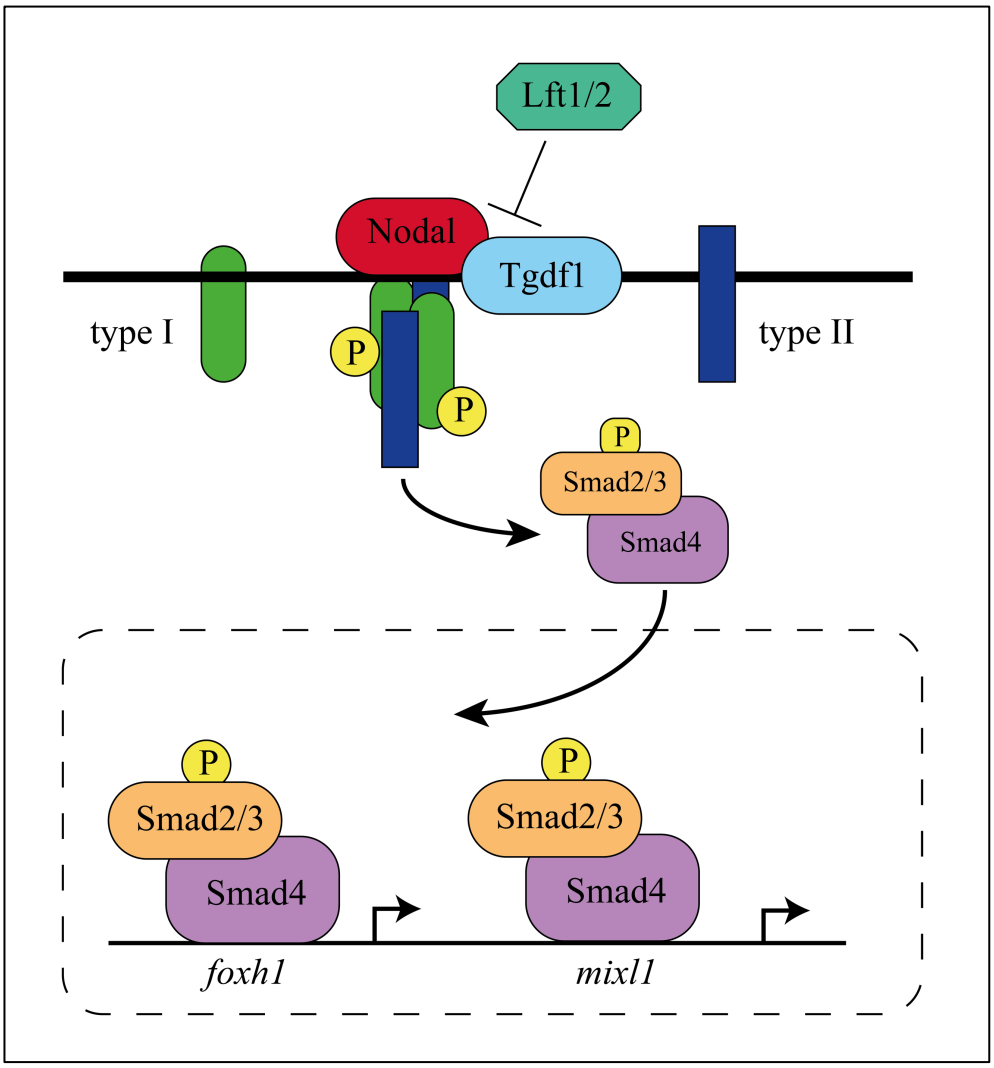


Figure 1.5 The Nodal pathway. Nodal signalling is transduced from the plasma membrane to the nucleus to regulate the transcription of target genes. The Nodal ligand binds to the type II receptor and the coreceptor Tdgf1 (transmembrane protein of the EGF-CFC family) activating the phosphorylation of the type I receptor. Smads2/3 are then phosphorylated, bind to Smad4 and the complex translocates to the nucleus where it binds to specific DNA-binding factors to allow transcriptional activation of specific targets such as *foxh1* and *mixl1*. Lefty 1 and Lefty 2 regulate extracellularly Nodal signalling by acting as ligand antagonists. Modified from Hill (2018).

A single *Nodal* gene is present in mice and humans and studies using mouse embryos revealed that total or partial loss of Nodal signalling results in truncation of the anterior mesendoderm (Hill, 2018) and at embryonic day E5.5, only the cells that ingress through the primitive streak express Nodal (Tam and Loebel, 2007). In pre-implantation mouse embryos, Nodal is also required for maintaining expression of genes determining pluripotency; for example, *Oct4/Pou5f1* and *Nanog* are sustained by Nodal signalling. Similar phenotypes and a loss of mesendodermal populations were observed in the absence of Activin A, which is

directly downstream of Nodal and binds to the same receptors as Nodal and initiates similar intracellular signalling events (Vallier et al., 2009).

Xenopus has six Nodal-related genes, *xnr-1, xnr-2*, *xnr-3*, *xnr-4*, *xnr-5* and *xnr-6* and with the exclusion on *xnr-3* are able to induce mesendodermal markers (Rex et al., 2002) while zebrafish have three *nodal* genes. Only *ndr1* (*nodal related protein 1*) and *ndr2* (*nodal related protein 2*), which are mutated in *squint* (*sqt*) and *cyclops* (*cyc*) mutants respectively, have a known role in endoderm formation. The third *nodal* related gene, *spaw (ndr3)*, is important for left-right patterning. *ndr1* is maternally expressed whilst *ndr2* and *spaw* are zygotic (Dougan et al., 2003; Long et al., 2003).

At the beginning of the blastula stage, *ndr1* and *ndr2* are expressed in a partially overlapping area in the vegetal marginal region of the blastoderm where the mesendoderm precursors are located. These precursors are under the control of another TF, Mxtx2, which is coordinated by Wnt signalling (Xu et al., 2012). Nodal ligands induce more *ndr1/2* expression in the blastoderm and the signalling domain keeps expanding; by the midblastula stage, *ndr1* is expressed in both the YSL and future mesendoderm, while *ndr2* is only found in mesendoderm, supporting the hypothesis that genes act on the progenitors to initiate the endoderm signalling cascade and then keep acting as a source on them (Bennett et al., 2007; Fan et al., 2007; Dubrulle et al., 2015; van Boxtel et al., 2015). Rigorous regulation of Nodal pathway activity is critical for the correct organisation of the mesendodermal cell population of the zebrafish gastrula, and whilst loss of function mutations in *ndr1* or *ndr2* result in mild defects in endoderm induction, embryos with null mutations in both *ndr1* and *ndr2* completely lack endoderm and trunk and head mesoderm (Feldman et al., 1998; Rebagliati et al., 1998; Sampath et al., 1998; Feldman et al., 2000; Chen and Schier, 2001; Dougan et al., 2003). This supports the idea that the two genes do act redundantly, and the combination of their signals converge to regulate the pathway (AND logic in a GRN network model). It is therefore the combination of *ndr1* and *ndr2* that leads to a specific outcome and activation of the Nodal signalling cascade. For example, if *ndr1* regulates genes X, Y and Z, whereas *ndr2* regulates genes X, Y and W; the lack of endodermal cells in the double mutant could be explained because Z and W are missing (thus the logic Ndr1 AND Ndr2). In the single mutant, both X and Y are present, leading to the existence of some endodermal cells. This proof of principle supports the idea that multiple genes work together to correctly fine tune a cell fate.

Predictably, overexpression of the Nodal antagonists Lft1 or Lft2 (Meno et al., 1999; Thisse et al., 2000) also leads to Nodal-deficiency phenotypes.

Not only is gene expression important but also the timing and duration of their expression in terms of cell fate decisions. This is evident if we consider that both *ndr1* and *ndr2* are essential for endoderm formation, nevertheless, not all the closest cells to the marginal zone become endoderm, with mesodermal cells also deriving from this region. This suggests that the underlying presence of additional regulators and signals is operating at these borders to determine endoderm and mesoderm (Warga and Nusslein-Volhard, 1999).

Studies have shown that a gradient of Nodal acts as a morphogen in endoderm specification; the cells closest to the margin at the midblastula stage are the ones exposed to the highest concentrations of Nodal and the majority of these cells take on an endodermal fate, whilst cells in tiers further away from the source domain (the YSL) are exposed to lower doses and therefore take on a mesodermal fate (Feldman et al., 2000; Thisse et al., 2000; Poulain et al., 2006; Dubrulle et al., 2015; van Boxtel et al., 2015; van Boxtel et al., 2018). The reason why some cells closest to the margin and therefore exposed to higher and longer durations of Nodal signalling become mesoderm instead of endoderm is yet not known. Ndr1 has been proven to function as a morphogen during mesoderm formation whereas Ndr2 does not appear to share this characteristic (Chen and Schier, 2001). Recent studies have expanded on this mechanism and shown how the domain of Nodal signalling evolves in space and time during embryonic development, introducing the concept of a 'temporal competence window' for Nodal signalling, as summarised in Figure 1.6 in the case of mesendoderm specification in the zebrafish embryo. A temporal window for competence of the Nodal pathway at the margin arises from the interplay between the Nodal ligands Ndr1 and Ndr2, the Nodal inhibitors Lft1, Lft2 and microRNA miR-430. This explains how the domain of Nodal activity is established and how it evolves over time (van Boxtel et al., 2015).

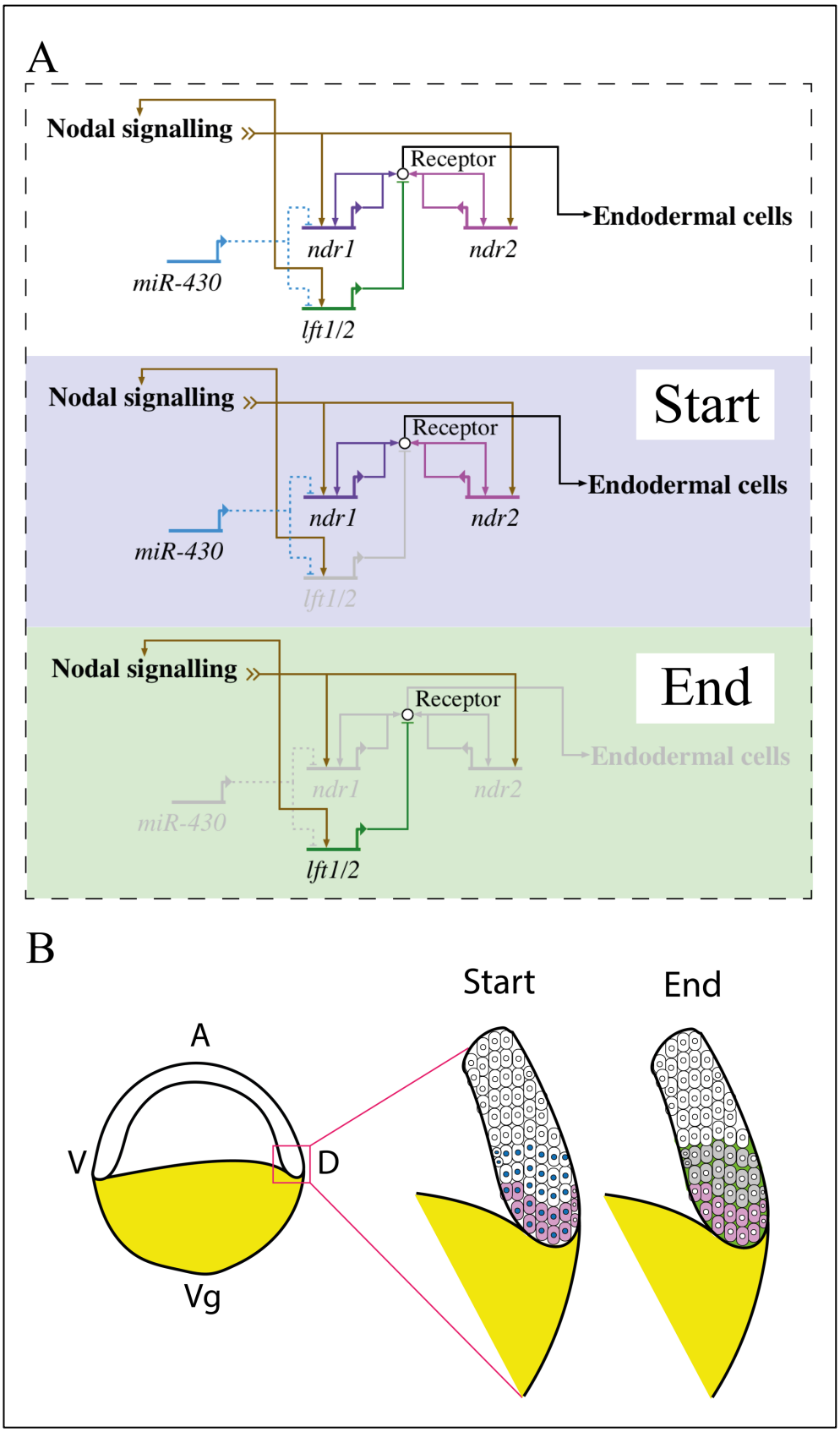


Figure 1.6 Evolution of the Nodal signalling GRN in the zebrafish embryo. (A) Overview of the GRN (Top panel). At the beginning (Start panel), Nodal signalling, in the cells at the margin, activates the expression of two Nodal ligands (Ndr1 and 2) and two Nodal antagonists (Lft1 and 2). The cells which are closest to the

YSL are exposed to higher concentrations of Nodal, however miR-430 delays the translation of the pathway inhibitors *lft1*/2 and zygotic *ndr1* by binding the 3'-UTR of these mRNAs. Binding of the miRNA to its target mRNA represses translation of the target protein without affecting the mRNA pool build-up. Maternal Ndr1 and zygotic Ndr2 accumulate in these cells and diffuse through the first 4 tiers of cells which are then channelled to endodermal fate. Note that miR-430 also inhibits *ndr1*, however the Nodal signal is more intense and overrides the microRNA inhibition. By 5.25 hpf (End panel), miR430 is degraded and translation of *lft1/2* is activated in the cells more distant from the YSL. Open circles represent extracellular interactions with Lft1/2 antagonists inhibiting the signal at the level of the Nodal receptor whilst Ndr1/2 activate the Nodal receptor directing endodermal specification. Lines ending in an arrow indicate positive activity, lines ending in a bar indicate inhibition. Dotted lines indicate translational repression. Greyed out text/lines indicate inactive gene/regulation. (**B**) Schematic representation of the Nodal domain at 5.25 hp. YSL cells expressing Ndr1/2 are shown in yellow; cells responding to Nodal are in pink (as the GRN in the Start panel). Cells expressing miR-430 have blue nuclei, grey shading represents cells where Nodal signalling is inhibited (higher distance from the YSL). Lft1/2 are translated when the inhibition by miR430 is overcome. Adapted from van Boxtel et al. (2015).

Nodal regulates Ndr1 and Ndr2 and their antagonists, Lefty1 (Lft1) and Lefty2 (Lft2) (Figure 1.6A panel start). In addition, Ndr1/2 positively regulates the expression of Lft1/2. Once Lefty proteins are translated, they block further Nodal signalling (Figure 1.6A panel end). The 'competency window' which dictates the spread/width of the Nodal signalling territory is regulated by the activity of a micro-RNA, miR-430 which delays the translation of Lft1 and Lft2. By 5.25 hpf, miR-430 is no longer able to suppress the levels of *lft1/2* mRNA from being translated and they start to inhibit Nodal signalling which has spread up to about five cell tiers from the YSL by this timepoint. It is the accumulation of these *lft1* and *lft2* transcripts that is mediated by miR-430 that allows a delayed response at first followed by a rapid induction in Nodal target genes, within this temporal window. In summary, different Nodal concentration thresholds induce different gene expression patterns and these findings suggest that both timing and transcription rate are important in determining the appropriate response to Nodal (van Boxtel et al., 2015).

As mentioned earlier in zebrafish, Ndr1 and Ndr2 act via a TGF-β receptor called Acvr1ba which is expressed in the same domain, the blastoderm margin (Aoki et al., 2002b). Overexpression of a constitutively active form of the Acvr1ba receptor upregulates endodermal gene expression throughout the zebrafish embryo (Alexander et al., 1999) and can cell-autonomously convert embryonic cells to an endodermal fate (David and Rosa, 2001). Embryos treated with inhibitors of type I or type II Nodal receptors, such as Antivin, fail to develop endoderm structures because the cascade of Nodal signalling is not activated (Thisse

et al., 2000). Another important factor for endoderm formation in zebrafish is Tdgf1 (Oep, One-eyed pinhead), an EGF-CFC membrane-associated protein (Schier et al., 1997). This extracellular coreceptor is required by Nodal ligands to bind to and activate their receptors and is expressed both maternally and zygotically. A phenotype similar to *ndr1;ndr2* (*sqt;cyc*) double mutants, with a lack of endoderm and most of the mesoderm (with the exception of some posterior mesoderm), is observed in embryos missing both maternal and zygotic Tdgf1 (Gritsman et al., 1999). The combination of these data supports the role of Acvr1ba as being sufficient and essential for endoderm formation and also confirm that Tdgf1 is necessary for Ndr1 and Ndr2 to bind their receptors and therefore start the Nodal signalling cascade.

Evidence is also emerging that Vg1 (also known as Gdf3) forms heterodimers with Nodal to instruct cells to become endoderm and mesoderm. It is the Nodal-Vg1 complex that activates the Nodal receptors; M*Zgdf3* mutants fail to form mesodermal and endodermal tissues (Bisgrove et al., 2017; Montague and Schier, 2017). Co-immunoprecipitation assays have confirmed the interaction between Nodal and Gdf family members (Fuerer et al., 2014), however in zebrafish it is unclear which heterodimer of Gdf3/Nodal is formed, i.e. whether it is Gdf3/Ndr1 and/or Gdf3/Ndr2 (or a combination of the two) that acts as an inducer of mesendoderm.

1.6 Other important signalling in endoderm formation

In addition to Nodal signalling, bone morphogenetic protein (Bmp) and fibroblast growth factor (Fgf) play important roles in determining the spatial and temporal expression of endoderm-specific genes. Starting from the blastula stage, Nodal ligands are expressed at the margin and YSL but on the ventral side an opposing gradient of Bmp (another TGF-β family member) suppresses endoderm specification. Poulain et al. (2006) showed how overexpression of a combination of *bmp2*, *bmp4* and *bmp7* signals led to reduced endoderm formation on the ventral side and their knockdown increases the number of endodermal cells. On the dorsal side, Nodal signalling is counteracted by Fgf signalling and Mizoguchi et al. (2006) showed that activation of the Fgf pathway caused a decreased number of endodermal cells expressing *sox32*. Equally, inhibition of Fgf signalling promotes endoderm development with an increase in the number of endodermal cells, without affecting the expression of Nodal.

In conclusion, initially, the formation of both endodermal and mesodermal progenitors relies on different doses and durations of Nodal signalling activity, however the end fate of cells is

also impacted by the differential induction and activity of other extracellular signals such as Bmps and Fgfs, which play repressive roles. How these factors all act intracellularly in combination to generate and maintain the proper ratio between endoderm and mesoderm is unclear.

1.7 Transcriptional control of endoderm formation

Genetic studies have shown that there must be additional factors involved at the beginning of the transduction cascade leading to the formation of endoderm and have started to highlight what they are. The Nodal signalling pathway induces the expression of TFs that appear to act in parallel during endoderm specification; these factors in chronological order of discovery include Mixl1 (also known as Bonnie and Clyde, Bon, Mixer) (Kikuchi et al., 2000), Gata5 (Faust, Fau) (Reiter et al., 1999; Reiter et al., 2001) and Sebox (Mezzo/Ogx9) (Poulain and Lepage, 2002). Downstream of Nodal ligands, the activation of these three genes is required for the expression of the *sox32* gene (Casanova, Cas) (Alexander et al., 1999; Dickmeis et al., 2001; Kikuchi et al., 2001). Pou5f/Spg is also required for *sox32* expression (Reim et al., 2004). Sox32 in turn induces the expression of *sox17*, the effects of Sox17 on the development of organs of endodermal derivation remains unknown. Similar to other vertebrates, the genes of the *foxa* family are also expressed in zebrafish endoderm during gastrulation and organogenesis (*foxa*/*fkd4*, *foxa1*/*fkd7*, *foxa2*/*fkh1*/*axial*/*hnf3b*, and *foxa3/fkh2*) (Odenthal and Nusslein-Volhard, 1998). Hhex TF has also been shown to be involved in the differentiation of some endoderm-derived organs, such as the liver, thyroid and pancreas (Wallace et al., 2001; Elsalini et al., 2003; Bischof and Driever, 2004; Gao et al., 2018). I will now describe each individual component of the network and the corresponding family in detail and reflect on their roles in other model species.

As introduced earlier, it is the combination of TFs and their redundant/overlapping roles that determines the acquisition of a specific cell fate. Nevertheless, some TFs are more important than others in initiating patterns of gene expression that result in major developmental changes. In endoderm formation, a critical role belongs to Sox32, which is both necessary and sufficient to induce endoderm (Alexander et al., 1999; Dickmeis et al., 2001; Kikuchi et al., 2001; Sakaguchi et al., 2001). It is found only in teleosts due to a duplication of Sox17 (Voldoire et al., 2017), orthologs of which in *Xenopus* and mouse are regarded as master

regulators of endoderm specification (Clements and Woodland, 2000; Kanai-Azuma et al., 2002; Clements et al., 2003; Sinner et al., 2004; Niakan et al., 2010).

Expression patterns observed by *in situ* hybridization have elucidated the spatial and temporal domain of these fundamental endodermal factors. Most of the TFs are expressed in the marginal cells of the zebrafish blastula (Figure 1.7), under the control of Nodal signalling and the opposing Fgf gradient. By 5.25 hpf epiboly the mesendodermal population is characterised by the expression of multiple markers scattered throughout the margin, with cell expressing multiple TFs at the same time. However, the distribution of these TFs along the animal-vegetal (and dorsal ventral) axis differs, consistent with the endodermal progenitors being located within the most marginal tiers and the mesodermal progenitors being present in both the most marginal tiers (intermingled with the endodermal cells) and tiers of cells located higher up (Figure 1.7B).

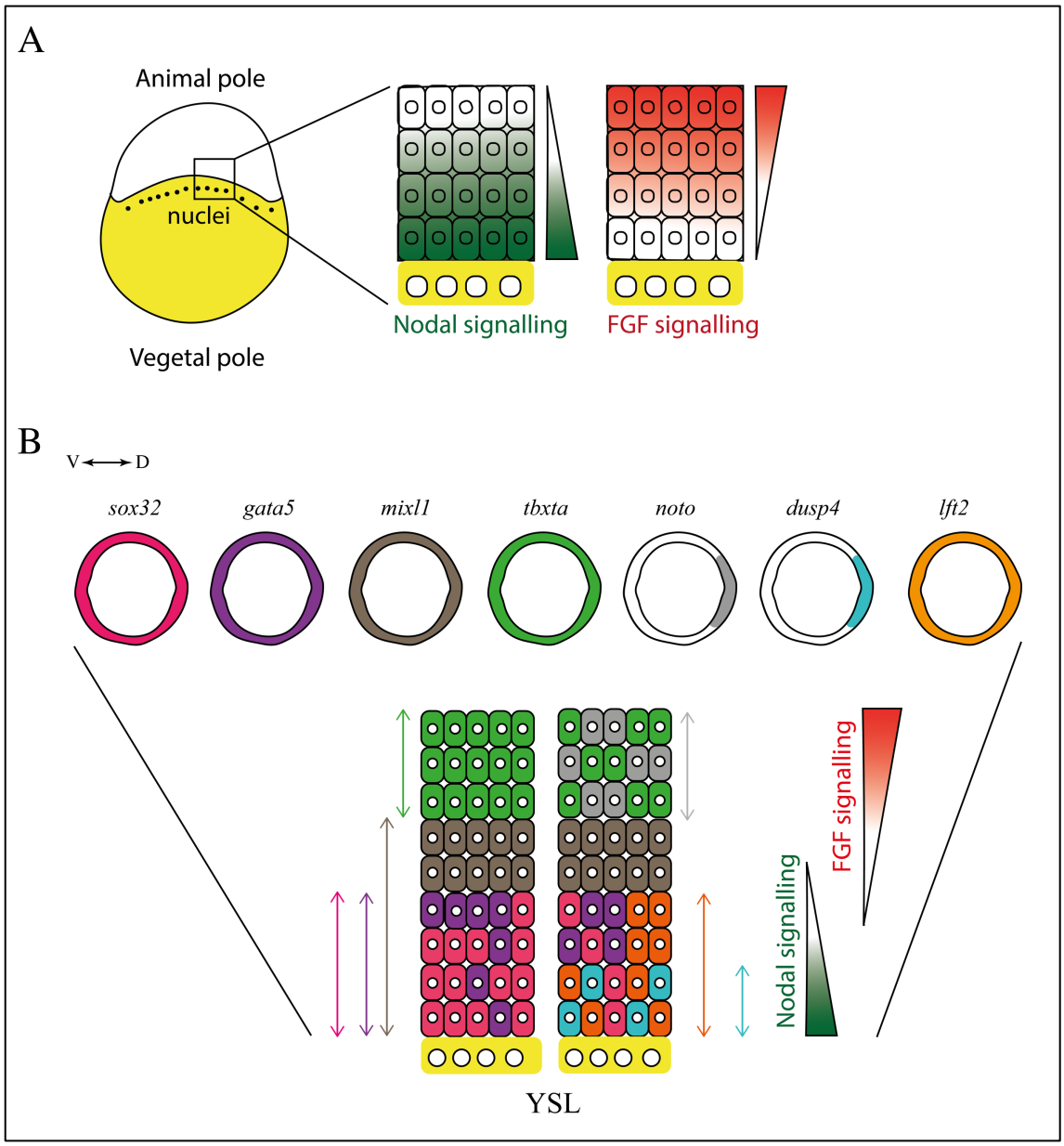


Figure 1.7 Combinatorial Nodal and Fgf signalling and a hierarchy of multiple TF modulate endoderm specification. (**A**) Schematic representation of Nodal and Fgf signalling pathways. An opposite gradient of Fgf signalling counteracts the Nodal signalling (high level at the YSL and low level at the animal pole) as shown at midblastula stage. The opposing activity of these two signal transduction pathways is important in organising the endodermal mesodermal boundaries. (**B**) Spatial domains of TFs expressed in endodermal and mesodermal precursors represented by the colours as shown. Note that the first 4-tiers are endodermal cells (expressing *sox32* and *gata5*) although some mesodermal cells are also present in these first 4-tiers. In the higher tiers of cells where Nodal signalling decreases, Fgf signalling is higher meaning that cells switch to a mesodermal fate (*tbxta* and *noto* expression). *lefty* expression is important to prevent Nodal signalling as previously described, whereas *dusp4* is important to inhibit the Fgf signal in the first 2-tiers by preventing phosphorylation of the Fgf downstream effector ERK.

1.7.1 The prominence of Sox factors in zebrafish endoderm formation

Sox family members Sox32 and Sox17 have both been shown to be important TFs involved in endoderm specification. Sox17 is present in mouse, human, *Xenopus* and zebrafish, whilst Sox32 is unique to the teleosts, where it appears that it takes on part of the role covered by Sox17 in other vertebrates. Although Sox17 is evolutionarily conserved in sequence from zebrafish to mouse, its functional role in endoderm formation is not necessary in zebrafish, where Sox32 is the required Sox family TF to establish the endodermal lineage (Alexander et al., 1999; Sakaguchi et al., 2001). In zebrafish, Sox17 is not essential for endoderm formation, indeed its function and the molecular events controlled by this gene are still largely unknown, whereas Sox32 plays an essential cell-autonomous role in endoderm formation (Dickmeis et al., 2001; Kikuchi et al., 2001; Aoki et al., 2002a).

As depicted in a simplified manner in Figure 1.8, *sox32* expression is first detected at midblastula (4.00 hpf) at the margin, the presumptive territory of the mesendodermal population (Figure 1.8A). Cross-sections of embryos show that during gastrulation, *sox32*-expressing cells involute at the margin, migrating around the YSL and spreading over the whole embryo in a scattered pattern (Melby et al., 1996; Alexander and Stainier, 1999; Pézeron et al., 2008). This embryonic germ layer corresponds to blastoderm cells mostly located in the dorsal half of the zebrafish at the midblastula margin, subjected to inductive signals (Nodal molecules) emanating from the YSL and the margin itself (Rodaway et al., 1999; Warga and Nusslein-Volhard, 1999).

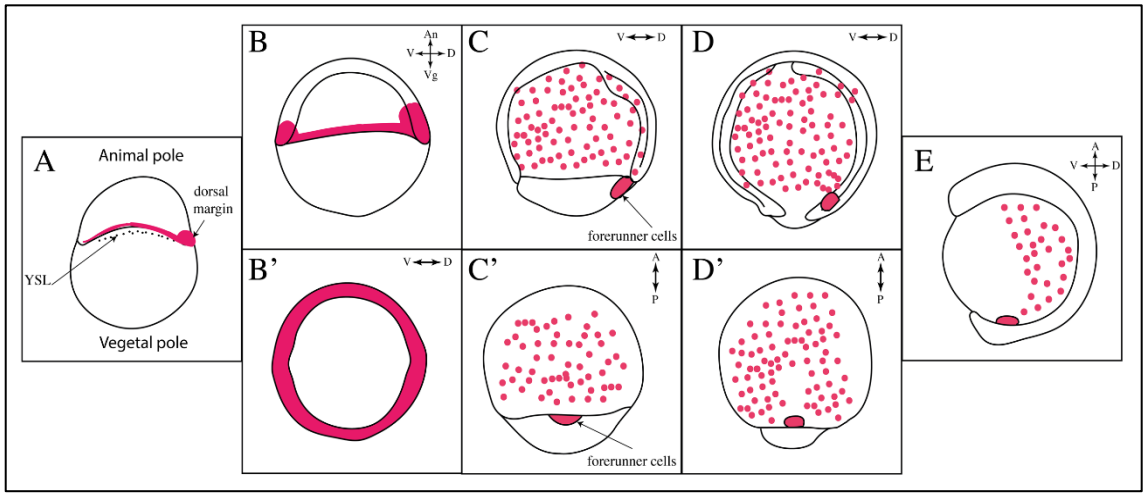


Figure 1.8 *sox32* expression patterns, between blastula and gastrula. (**A**) At blastula stage, endoderm cells are located at the margin. Nodal signalling from the YSL directly induces the expression of *sox32*. (**B**) and (**B'**) At 5.25 hpf, *sox32* is still expressed in the YSL and in the marginal cells, mostly on the dorsal side.

(**C**) and (**C'**) *sox32*-expressing cells have begun to involute and by 7.00 hpf *sox32* is expressed in scattered endodermal cells in a recognisable 'salt and pepper pattern'. Expression is also observed in the forerunner cells (arrows). (**D**) and (**D'**) At the end of gastrulation (9.00 hpf), *sox32* endodermal cells have been internalised and by the (**E**) 1-2 somites stage (10-11 hpf) they have formed a monolayer around the developing gut. Embryos in A, B, C, D and E are in lateral view with dorsal to the right; B' is animal view of B. C' and D' are dorsal view, anterior to the top, of C and D.

Nodal signalling indirectly controls *sox32* expression through the regulation of a set of TFs including Mixl1, Eomes, Gata5 and Sebox which in turn directly activate *sox32* (Reiter et al., 1999; Kikuchi et al., 2000; Reiter et al., 2001; Pereira et al., 2012; Nelson et al., 2014; Xu et al., 2014). Sox32 mutants display absence of the gut tube and related endodermal organs due to the fact that markers of endoderm differentiation such as mixl1 and sox17 at the dorsal margin at the onset of gastrulation are not expressed (Alexander et al., 1999). Furthermore, due to the reciprocal endoderm-mesoderm interactions, proper development of the zebrafish cardiovascular system is impaired and Sox32 mutants develop cardia bifida, the formation of bilateral hearts (Alexander et al., 1999; Sakaguchi et al., 2001). Ectopic expression of *sox32* is sufficient to convert mesodermal precursors into endoderm in the dorsal marginal zone during the late blastula stage and in embryos lacking *sox32* activity, cells that are normally fated to become endoderm adopt a mesodermal fate. However, the conversion is restricted to the margin area, and has no effect on ectodermal cells, indicating Sox32 interacts with other factors that are only present at the margin (Aoki et al., 2002a).

In the absence of Nodal signalling in the MZ*tdgf1*, where *sox32* expression is diminished in addition to *mixl1* and *gata5*, injection of *sox32* mRNA is sufficient to restore *sox17* expression and endoderm formation. This same approach also works to restore endoderm formation in Mixl1 and Gata5 mutants. These experiments place *sox32* downstream of these transcriptional regulators in endoderm formation. On the other hand, it has been shown that *sox32* operates upstream of *sox17*, because the mutation or knockdown of the *sox32* gene completely abolishes *sox17* expression, leading to the loss of endodermal cells (Alexander et al., 1999; Kikuchi et al., 2001). Sox32 has also been shown to interact with Pou5f3 (also known as Pou5f1 or Oct4) to regulate its own expression and to bind to the *sox17* promoter to promote endoderm development (Lunde et al., 2004; Chan et al., 2009a; Perez-Camps et al., 2016).

Pou5f1 encodes a POU domain TF which is both maternally and zygotically expressed (Lunde et al., 2004; Lee et al., 2013; Lippok et al., 2014; Perez-Camps et al., 2016; Voronina

and Pshennikova, 2016). Analysis of the *pou5f3* mutant MZ*spg* (*spiel-ohne-grenzen*) which lacks both maternal and zygotic Pou5f3 (Reim et al., 2004; Voronina and Pshennikova, 2016) showed defects in mesoderm and endoderm formation and demonstrated that Pou5f3 is necessary to maintain *sox32* expression (Lunde et al., 2004; Reim et al., 2004). In the MZ*spg* mutant, *sox32* is detected at the blastula stage in mesendodermal cells, however because zygotic Pou5f3 protein is not functional, Sox32 is not maintained during the gastrula stage. As a result, the proper ratio between mesodermal and endodermal fates is disrupted (Dickmeis et al., 2001). As *sox32* is both necessary and sufficient for endoderm formation, it is hence suggested that *sox32* functions as the key cell fate regulator that controls the endoderm lineage.

As mentioned earlier, Sox32 acts upstream of *sox17* and plays a critical role in early endoderm formation in zebrafish. In mouse, SOX17 is expressed in both visceral and definitive endoderm and in conjunction with GATA6, is a master regulator of endodermal cell fate (Artus et al., 2011). Sox17 deletion not only causes gut defects but is lethal in embryonic mice (Kanai-Azuma et al., 2002). Similarly, Sox17 contributes to endoderm formation in *Xenopus* (Hudson et al., 1997). In zebrafish, *sox17* transcript expression is first detected at the midblastula stage in the cells at the dorsal margin and is later confined to the endodermal progenitors (Alexander and Stainier, 1999). During gastrulation, *sox17* expression is seen within the migrating and involuting cells and the forerunner cells, a non-involuting cell population that will go on to form the Kupffer's vesicle. This fluid filled sac then eventually becomes part of the mesoderm as notochord and muscle. Throughout gastrulation, Sox17 marks migrating endoderm cells and its expression disappears after somitogenesis. The pattern of *sox17* expression through gastrulation is depicted in Figure 1.9.

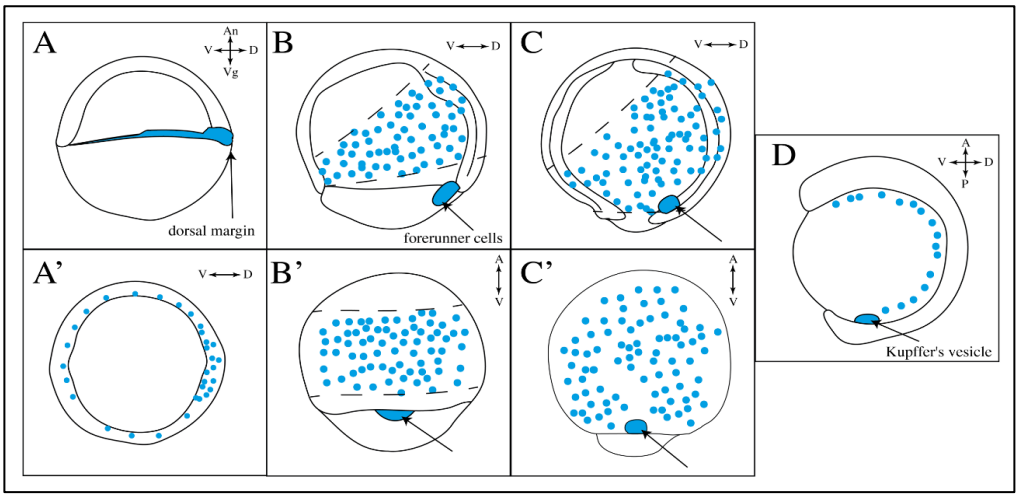


Figure 1.9 Dynamic expression of *sox17* during zebrafish development. (**A**) and (**A'**) *sox17* is first detected

before gastrulation in a dorsally located group of marginal cells and at the onset of gastrulation, *sox17* is found scattered throughout the margin. (**B, C**) and (**B', C'**) *sox17* is expressed in endodermal precursors throughout gastrulation; the migration of these cells gives a similar 'salt and pepper' pattern as seen for *sox32.* In addition, *sox17* labels the forerunner cells (arrows in B and C). (**D**) At the start of somitogenesis, *sox17* is expressed in a monolayer of cells that will give rise to the embryonic gut and pharyngeal arches. Arrow depicts Kupffer's vesicle which is formed from the forerunner cells. Embryos in A, B, C, D and E are in lateral view with dorsal to the right. A' is animal view of A. C' and D' are dorsal view, anterior to the top, of C and D. An: animal, Vg: vegetal, V: ventral, D: Dorsal, A: anterior, P: posterior.

Overexpression of *mixl1* increases the number of *sox17* expressing cells and promotes *sox17* expression in MZ*tdgf1* mutants suggesting that *sox17* is downstream of Mixl1 and Tdgf1. HMG binding domains exist in the *sox17* promoter, advocating that it is regulated by another Sox factor, specifically Sox32 (Ober et al., 2003). In support of this, overexpression of Acvr1ba promotes *sox17* expression in wild type and MZ*tdgf1* mutants but not in *sox32* mutants (Alexander and Stainier, 1999). Like *sox32*, *sox17* overexpression can repress mesodermal markers (Aoki et al., 2002a).

Chan et al. (2009) further clarified the importance of both Sox32 and Pou5f3 in synergistically activating *sox17* expression during endoderm specification, by actively binding the regulatory regions upstream of *sox17.* They further validated that the activation and subsequent sustained expression of *sox17* during gastrulation depends only partially on Nodal signalling. Sox32 is downstream of Nodal targets, however Pou5f3 expression is not. This introduces another layer of regulation and fine tuning in the endoderm pathway with Pou5f1 acting in parallel to Sox32.

1.7.2 Mix-like paired homeobox TFs are critical for endoderm development

Other factors which function in the early vertebrate embryo and are intrinsically connected to promotion of endoderm formation are Mix-like paired homeobox proteins. There are four Mix-related genes in zebrafish (*mixl1*, *sebox*, *mxtx1* and *mxtx2*) all of which are involved in endoderm formation (Kikuchi et al., 2000; Poulain and Lepage, 2002; Sakaguchi et al., 2006; Xu et al., 2012). Overall, Mixl1 mutants develop fewer endodermal cells, consistent with a reduction in the level of *sox32* expression in these mutants (Alexander et al., 1999; Kikuchi et al., 2000). At the onset of gastrulation, both *mixl1* and *sebox* are expressed in the mesendodermal progenitors at the margin like *sox32*; *mixl1* is then rapidly downregulated and becomes undetectable by 60% epiboly (Alexander et al., 1999). Similar to in *Xenopus*, Nodal

signalling initially induces *mixl1* expression in zebrafish, likely through binding phosphorylated Smad2 and regulates endodermal gene expression (Kikuchi et al., 2000; Poulain and Lepage, 2002). *mixl1* expression is abolished in *ndr1;ndr2* double mutants (Alexander and Stainier, 1999) suggesting that *mixl1* is downstream of the Ndr1/Ndr2 receptor. In conjunction with this, *mixl1* expression is lost when Nodal receptors are inhibited in antivin-injected embryos (Alexander and Stainier, 1999). The T-box TFs, Tbxta and Tbx16, which themselves are Nodal dependent genes, are needed for *mixl1* expression, further supporting the necessity of a continual Nodal signalling for endodermal and mesodermal precursors before gastrulation (Nelson et al., 2017). Sebox is also necessary for endoderm genes induction as shown by experiments with the Mixl mutant, where the associated phenotype is partially rescued by injecting Sebox mRNA (Poulain and Lepage, 2002). Moreover, when a Sebox MO is injected into the Mixl1 mutant, all endodermal gene expression is completely abolished, and severer defects are observed. Complementary to these data, overexpression of Sebox induces ectopic endodermal gene expression and reduces mesoderm, a similar pattern to that observed when *sox32* is overexpressed in the margin (Poulain and Lepage, 2002). It therefore follows that Sebox is necessary in zebrafish endoderm formation and that it works in parallel to Mixl1 and Gata5. It is still unknown how, and at what level of the endodermal pathway, Sebox is located, or if other partners are interacting with it. What is known however, is that it regulates its own expression and cooperates with Mixl1 and Gata5 to activate Sox32, either directly or indirectly, and to maintain downstream endoderm markers (Poulain and Lepage, 2002).

1.7.3 The role of Gata family TFs in patterning early endoderm

Other factors key for endoderm formation are Gata family members. The *gata* genes encode for TFs characterised by zinc finger DNA-binding domains that bind to the consensus DNA sequence (A/T)GATA(A/G) (Gronenborn, 2005). Their conserved roles in embryogenesis, both in endoderm and mesoderm formation, have been shown by knockdown experiments in worms (Zhu et al., 1998), flies (Rehorn et al., 1996), fish (Reiter et al., 1999; Tseng et al., 2011), frogs (Weber et al., 2000; Afouda et al., 2005) and mice (Decker et al., 2006; Rojas et al., 2010; Artus et al., 2011; Carrasco et al., 2012). During the gastrula stage of *Xenopus*, Gata4/5/6 acts downstream of Nodal signalling in endoderm formation (Afouda et al., 2005), and overexpression of Gata5 is sufficient to induce ectopic endodermal gene expression (Weber et al., 2000). In zebrafish, Gata TFs are separated into two subfamilies; the first being

comprised of Gata1, Gata2 and Gata3 which play a major role in haematopoiesis (Heicklen-Klein et al., 2005) and the second containing Gata4, Gata5 and Gata6, which are mainly involved in endoderm and heart formation (Reiter et al., 1999; Holtzinger and Evans, 2005, 2007). Among the second endodermal subfamily, *gata5* is the most characterised, as shown by the available *faust (fau)* mutant line. This mutant shows impaired endoderm formation with defects in gut morphogenesis that range from a lack of gut looping to severely reduced endoderm formation (Reiter et al., 1999; Reiter et al., 2001). No mutants for Gata4 and 6 are currently available. Similar to Mixl1, Nodal signalling directly regulates *gata5* expression before gastrulation in endodermal progenitor cells (Alexander and Stainier, 1999) which then initiate *sox32* (David and Rosa, 2001; Dickmeis et al., 2001; Reiter et al., 2001; Sakaguchi et al., 2001) which consequently promotes the expression of *sox17* and endoderm specification. Similar to other mesendodermal markers, *gata5* and *6* are found at the blastoderm margin and they continue to be coexpressed in the endoderm and YSL at 8 hpf. In the zebrafish embryo, *gata*5 and 6 are dynamically expressed in the developing mesendodermal and endodermal cells; their expression peaks first at 5 hpf and then again at 11 hpf when they are observed in the endoderm, ventral mesodermal derivatives and the heart primordium (Tseng et al., 2011). *gata5* expression at the midblastula stage is more marginally restricted when compared to *mixl1* expression (Reiter et al., 2001), whereas *gata4* and *6* are expressed in the posterior endoderm after gastrulation has begun (Reiter et al., 2001; Tseng et al., 2011). Similar to *sox17*, the expression of *gata5* decreases from 16 to 24 hpf; at the latter timepoint it becomes detectable again and all three TFs (*gata5, gata6* and *sox17*) are observed in the gut and endoderm-derived organs (liver, pancreas). *gata5* and *gata6* are also detectable in the heart at this timepoint (Molkentin, 2000). In medaka fish, *gata5* and *gata6* are also temporally coexpressed during endoderm formation (Kobayashi et al., 2006). Similar patterns of expression were observed in mouse, where GATA6 was also detected later in the development of endoderm derived tissues and is required in the initial specification of the pancreas (Decker et al., 2006; Carrasco et al., 2012). Additional proof of Gata factors' roles in endoderm formation comes from mammalian cell transfection studies (human and mouse) in which expression of *Gata4*, *Gata5*, or *Gata6* activates downstream endodermal genes (Holtzinger et al., 2010; Fisher et al., 2017; Yiangou et al., 2018; Chia et al., 2019).

1.7.4 Forkhead TFs play a critical role in endoderm formation

Forkhead factors are also expressed during vertebrate gastrulation in endoderm precursors during vertebrate gastrulation and additionally play an essential role during several stages of vertebrate mesoderm and endoderm formation and patterning (Dirksen and Jamrich, 1995; Alexander et al., 1999; Stainier, 2002; Grapin-Botton and Constam, 2007; Golson and Kaestner, 2016). They are members of the winged helix/forkhead family of nuclear TFs and in humans have been shown to be 'pioneer' factors (Strahle et al., 1993; Zaret and Carroll, 2011), opening the compacted chromatin for other proteins through interactions with nucleosomal core histones. Deficient (FOXA2$^{-/-}$) mouse embryos lack completely the foregut endoderm (Lee et al., 2005). In zebrafish, nine *fox* genes have been reported, three genes *foxa1*, *foxa2* and *foxa3* are expressed in endodermal cells in a sequential and overlapping pattern (Odenthal and Nusslein-Volhard, 1998). Of the 3 forkhead family members, Foxa2, formerly known as Axial or HNF3β, acts downstream of Nodal signalling and Sox32 and upstream of Sox17. *foxa2* expression starts at shield stage and like *sox32* and other mesendodermal markers is located in the margin cells (Alexander et al., 1999).

1.7.5 T-box TFs are important in activating expression of mesendodermal genes.

Another TF that is localized to marginal blastomeres in zebrafish and regulates Nodal signalling is the T-box TF Eomesodermin (Eomes). Various genetic screens have showed that Eomes is both necessary and sufficient for mesoderm induction in *Xenopus* (Vignali et al., 2000) and that Eomes plays a role in induction of both endoderm and cardiac mesoderm in mice (Nowotschin et al., 2013). Recent work with human ESCs and mouse epiblast stem cells has also established the importance of Eomes acting downstream of phosphorylated Smad2/3 in the specification of definitive endoderm, using a combination of whole-genome expression and ChIP-seq analyses (Teo et al., 2011). Two Eomesodermin homologues, *eomesa* and *eomesb* (Takizawa et al., 2007) have been identified in zebrafish, with Eomesa being a maternal determinant and instrumental, in combination with the TF FoxH1, in specifying mesendoderm (Slagle et al., 2011; Nelson et al., 2014) and in conjunction with Gata5 and Mixl1, Eomesa regulates the endodermal gene *sox32* (Bjornson et al., 2005). Overexpression of Eomesa in WT embryos leads to induction of dorsal mesodermal markers and knockdown of Eomesa by MO results in endoderm deficiency, confirming the dual role of the TF (Xu et al., 2014). Significantly, double morphants of *eomesa* and *mixl1* are characterised by low levels of endoderm (Bjornson et al., 2005). Overexpression of Eomesa in MZ*tdgf1* embryos,

which lack both maternal and zygotic copies of *tgdf1* and normally fail to develop endoderm and most mesodermal structures, induces *sox32* expression, bypassing endoderm formation by Nodal signalling linked to Tgdf1 receptor, and thus the resulting embryos present only slightly reduced expression of endoderm markers and moderate lethality by 24 hpf. On the other hand, Eomesa does not induce ectopic expression of *ndr1* and *ndr2* in MZ*tdgf1* mutants (Xu et al., 2014). Immunoprecipitation assays have confirmed that Eomes is able to bind to and pull down both Gata5 and Mixl1 TFs, reaffirming that the earliest marker of endoderm progenitors, Sox32, requires the cooperativity of multiple TFs for its induction. Reciprocally, Mixl1 and Gata5 are able to pull down Eomesa (Bjornson et al., 2005).

These experiments show how Eomesa works together with Mixl1 and Gata5 to initiate and maintain *sox32* expression in endodermal lineages. Importantly however, while Eomesa is sufficient for induction of mesoderm and endodermal genes it is not absolutely required for their expression. This supports the idea introduced earlier that multiple signalling pathways and factors function combinatorially to form and pattern the endoderm germ layer, by interacting and utilising different pathway components to perform redundant roles.

1.8 GRN of endoderm development

Two research papers have visually summarised our knowledge to date of endoderm formation in zebrafish (Ober et al., 2003; Slagle et al., 2011), however, no updated version has been published since 2011 (Figure 1.10A). Chan et al. (2009) published the first systematic attempt to provide an integrated model of the GRNs underpinning zebrafish development using BioTapestry software, but the integrative model has not been updated since (Figure 1.10B). BioTapestry is an interactive tool for building, visualising and modelling GRNs that allows the researcher to construct a network model and use it to visualise and understand the dynamic behaviour of a complex, spatially and temporally distributed GRN (Longabaugh, 2012). This provides the researcher with the ability to break down, simplify and unravel a multifaceted pattern in order to study the complexity of gene regulation related to development; inferring regulatory interactions and identifying functionally related genes showing patterns of coexpression.

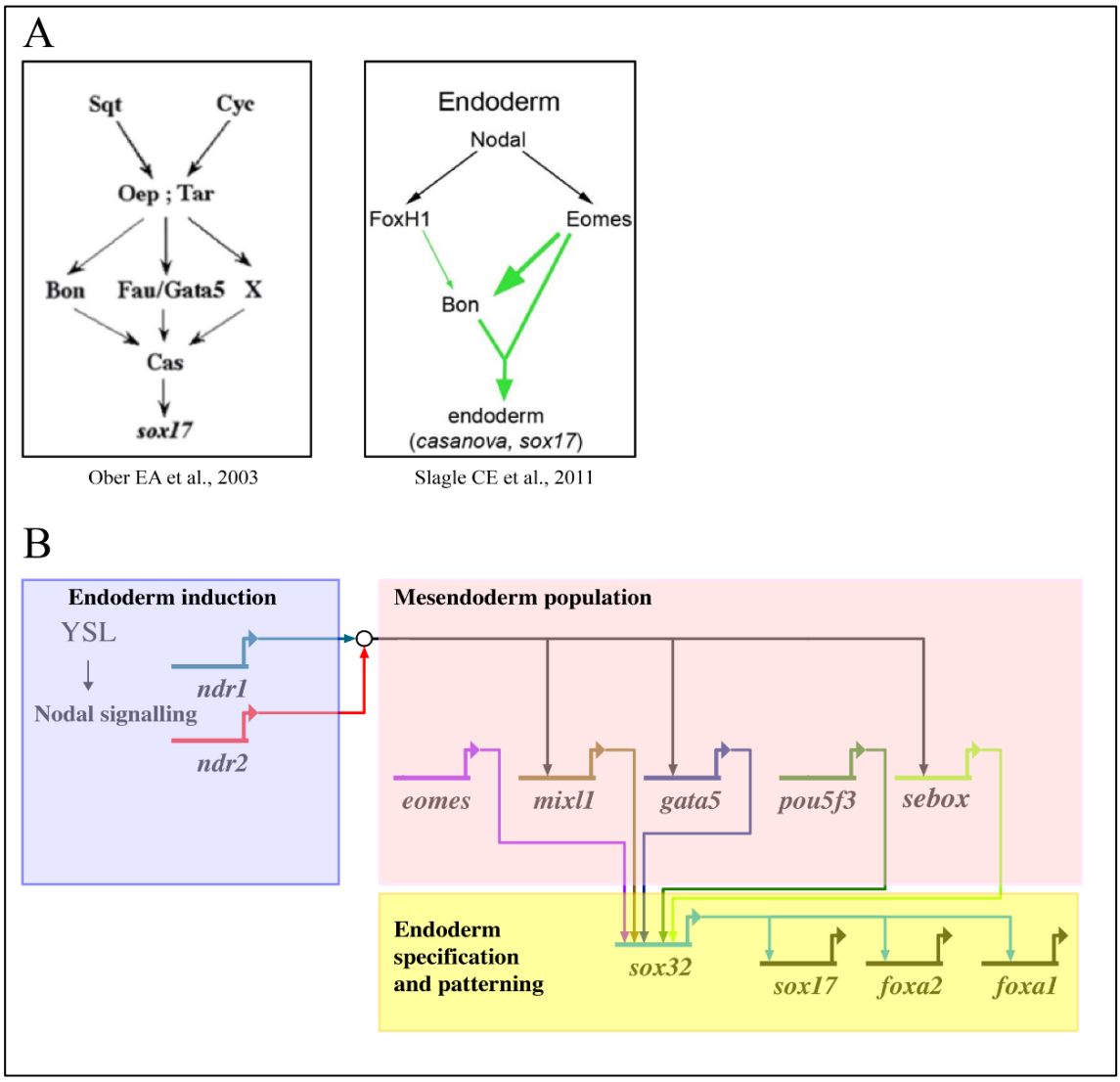


Figure 1.10 Signalling proteins and TFs that function within the elucidated pathways of zebrafish endoderm development. (**A**) Schematic representation of the signaling cascades involved in endoderm induction, specification and regionalisation during zebrafish development as proposed by two research papers (see text for full details). (**B**) BioTapestry was used to create a View from the Genome for the core genes driving endoderm formation and their epistatic interactions from the aforementioned reviews in (A). The shaded areas represent different spatial and temporal regions of expression: YSL (purple) induction at midblastula stage; mesendodermal population in the dorsal margin of the embryos (pink) and endodermal cells (yellow) during gastrulation. Lines ending in an arrow indicate positive interactions. All the epistatic relationships between nodes have been published but not in all cases has a direct interaction been validated.

In the endoderm model, the activator signal is Nodal originating from the YSL, with the intercellular ligands *ndr1* and *ndr2* activating the intracellular phosphorylation of Smad factors to upregulate several TFs (*eomes*, *mixl1*, *gata5* and *sebox*). These input factors regulate *sox32* which has been shown previously to activate the expression of *sox17* directly (Chan et al., 2009a) and the expression of *foxa1* and *foxa2* (Alexander and Stainier, 1999). In addition,

experiments show that Pou5f3 and Sox32 bind together to form complexes to specify mesendoderm cells (Lunde et al., 2004; Reim et al., 2004; Perez-Camps et al., 2016).

Other roles of Sox32 in the regulation of the network are poorly understood. Other key questions yet to be answered include: what are the negative feedback loops in the mesendodermal population? Which other factors have transient roles and orchestrate or restrict endodermal specification fate? Is the role of these other factors sequential, parallel or interconnected to the known TFs? How do the activities of Nodal and other important early developmental signaling pathways such as Bmp, Fgf, Ra, Wnt and Hh lead to the definition of the endodermal population?

There have been many studies of zebrafish development, however there is a lack of an updated global overview of the dynamics of the genetic architecture during germ layers formation. Similarly, to my knowledge, there has been no integrative study of the regulatory circuitry of lineage bifurcations since 2009, prompting me to undertake a systematic study of endodermal TF as a key objective of my thesis. The lack of both primary papers and reviews summarising this information is striking when compared to the GRNs available for other species. The mesendoderm and ectodermal GRN in *Xenopus* (Charney et al., 2017b; Maharana and Schlosser, 2018), the mesendodermal GRN in sea urchin (Peter and Davidson, 2010; Rafiq et al., 2014; Erkenbrack et al., 2018), GRN for specification of endomesodermal lineages in sea squirt (Shi et al., 2005; Kubo et al., 2010; Satou and Imai, 2015), the regulatory network repository of TFs for humans and mice (Mostafavi et al., 2014; Parfitt and Shen, 2014; Herpin and Schartl, 2015; Liu et al., 2015; Yun et al., 2017) are all available, and there is a plethora of information on the differentiation of stem cells (Zhou et al., 2007; Mohammadnia et al., 2016; Chia et al., 2019). In zebrafish, only certain specific GRNs have been developed; for example, Morley et al. (2009) exposed the role of Tbxta in the mesodermal network. Similarly, Greenhill et al. (2011) used a systems biology approach to identify and develop the GRN underlying melanocyte specification and differentiation while an overview of the GRN that orchestrates the formation of neural crest cells has been described by Petratou et al. (2018) and Williams et al. (2018). Most recently, Nelson et al. (2017) provided new information about the genes that are important in establishing mesendodermal identity in the early zebrafish gastrula. One of the key aims of my PhD was to collect and utilise these recent findings, and others, to update the zebrafish mesendoderm GRN from the beginning of gastrulation through to the end, linking together critical signalling pathways with their transcriptional targets and

most importantly, highlighting new regulatory connections related to endoderm specification (Figure 1.10B yellow quadrant).

1.9 Previous work leading to the project and aims

Recent applications of high throughput analysis such as microarray and NGS have helped close the gap in our understanding of genetic programmes during the development of model organisms. The field of 'omics' research has contributed greatly to large-scale genome-wide association studies and more recently, an increasing number of zebrafish laboratories have adapted genomics technology to investigate several aspects of zebrafish biology, from understanding the regulation of gene expression to epigenetic modifications of the genome.

An evolutionarily conserved molecular pathway that specifies endoderm during vertebrate gastrulation, including the activity of a number of TFs and signalling families, has been identified using forward and reverse genetic approaches since the end of the last century. However, these approaches are time consuming, expensive and labour intensive and are not suitable for the identification of important missing key nodes in the endodermal GRN. RNA-seq, ribosomal profiling, ChIP-seq for histone modifications and ChIP-seq for sequence-specific TFs have all started to dissect gene regulatory logic in zebrafish development from a genome-wide prospective, particularly in respect of the epigenetic landscape during zebrafish development has been studied with a focus on the midblastula transition (MBT), zygotic genome activation (ZGA), mesodermal and ectodermal lineages. However, the same technologies have yet to be applied to directly delve into the specific GRN of endoderm formation.

My initial aim was to further investigate the role of Sox32, Sox17 and Mixl1 in endoderm development in zebrafish embryos to identify direct and indirect targets by ChIP-exo. The HMG box TF Sox32 is a key component of the endodermal pathway and is essential for its formation; however, the molecular events controlled by Sox32 are largely unknown. In addition, Sox17, which is regarded as the master regulator of endoderm in *Xenopus* and mouse, seems to have been relegated to a secondary role in zebrafish endoderm, but how and why? It is not clear whether Sox32 and Sox17 preferentially target the same endodermal cascade of genes, or whether these TFs are linked to a niche specialisation in regulatory function. It is also unclear, how the Mixl1 TF, which is absolutely required in zebrafish for endoderm formation upstream of *sox32*, regulates expression of endodermal genes.

My next aim was to investigate endodermal cell lineage specification, focusing on the transcriptional circuitry necessary to generate/underpin endoderm formation. To address this question, I analysed the transcriptome in both Sox32 and Mixl1 mutants, and then compared these to the wild type. To enrich these data, I also implemented FACS-seq (fluorescence-activated cell sorting of cell populations followed by RNA-seq) to sort out and sequence populations of endodermal cells from the *sox17:GFP* transgenic line in order to provide a clearer picture of the endoderm specific transcriptome and to better understand gene expression in endodermal cells.

My third and final aim was to generate a more comprehensive GRN underpinning endodermal fate in zebrafish embryos by combining the data I generated with existing published information. In doing so, I intended to provide an updated map of how cells become committed to an endodermal fate in the zebrafish embryo.

The findings presented here extend our current understanding of endoderm formation in zebrafish embryos by the identification of novel genes involved in endoderm specification and the characterisation of gene regulation by Sox32, Mixl1 and other factors.

Chapter 2 – Material and Methods

Zebrafish work and husbandry

2.1.1 Zebrafish breeding and embryo handling

All experiments were carried out in accordance with Home Office recommendations and all animal procedures were performed under license as required by the Animals (Scientific Procedures) Act 1986 (UK). Zebrafish (*Danio rerio*) were raised at 26°C on a 14/10 hour light/dark cycle at the King's College London aquatics facility. Wild type AB strain zebrafish were used for chromatin immunoprecipitation (ChIP) assay validation; whilst the Tubingen strain was used for RNA-seq experiments. Zebrafish are photoperiodic in their breeding, mating and spawning in the morning after the sunrise. Hence, breeding behaviour was induced by the onset of artificial light and the embryos were collected with a net and placed in a petri dish. Unfertilized eggs were discarded; fertilized eggs were incubated at either 28.5°C or 33°C in petri dishes with a density of ≤200 embryos/dish, until the desired developmental stage was reached. Embryos older than 24 hours, used for *in situ* hybridization were treated with 1-phenyl 2-thiourea (PTU, Sigma-Aldrich, Cat#P7629), final concentration 0.003% to prevent pigmentation. Developmental stages were classified according to morphological features corresponding to respective age in hours post fertilisation (hpf) (Kimmel et al., 1995) using a dissecting microscope.

2.1.2 Zebrafish lines

Two mutant lines, *mixl1*^{m425} (Kikuchi et al., 2000; Stainier et al., 1996) and *sox32*^{ta56} (Chen et al., 1996; Solnica-Krezel et al., 1996), and one transgenic line *tg(sox17:GFP)* (Chung and Stainier, 2008) were used. The homozygous *mixl1* mutant (*mixl1*$^{-/-}$) fish were a kind gift from Prof. Dirk Meyer and Dr. Patrick Fischer, University of Innsbruck aquarium. Carriers of the *sox32*^{ta56} allele (*sox32*$^{+/-}$) were purchased from the Zebrafish International Resource Center (ZIRC) and maintained on the AB background. Embryos obtained from crosses between *sox32*$^{+/-}$ mutants were screened for homozygosity (*sox32*$^{-/-}$) as detailed below, while transgenic embryos generated from *Tg(sox17:GFP*) in-crosses were identified by screening for GFP expression. Heterozygous *tg(sox17:GFP)* carriers were outcrossed to wild type (AB

background) fish to maintain the line, while the homozygous fish were used in flow cytometry experiments (below).

2.1.3 *sox32* mutant genotyping

The *sox32* allele is a thymidine to cysteine (C-T) mutation at nucleotide 510 of the second coding exon of *sox32*. This mutation is predicted to generate a premature stop codon at residue 170 of the protein, causing truncation of the protein shortly after the HMG-box (high mobility group box) domain (Alexander et al., 1999). In order to genotype individual fish, caudal (tail) fin clipping was performed on 3 month old zebrafish and genomic DNA extracted from the tissue using the HotSHOT method (Meeker et al., 2007). Briefly, fish were anaesthetised by immersion in 0.02% MS-222 (Ethyl 3-aminobenzoate methanesulfonate, Sigma-Aldrich Cat#E10521) until gill movement was reduced and swimming ceased. The anesthetised fish was held using a plastic spoon whilst approximately half of the caudal fin was clipped using clean, surgical scissors. Fish were then immediately transferred to individual holding tanks containing fresh aquarium system water with 0.1% methylene blue and monitored until they recovered and recommenced swimming. This procedure lasted less than two minutes, with regrowth of the clipped fin occurring in approximately two weeks (Azevedo et al., 2011). The tissue samples obtained were suspended in 40µl of lysis buffer (50mM NaOH, 0.2mM EDTA) and incubated at 95°C for 20-40 minutes with vortexing, then cooled to 4°C after which 10µl of neutralisation buffer (1M Tris-HCl pH 8) was added. The volume was then adjusted to 100µl with double distilled water (ddH_20) and cell debris pelleted by centrifugation at 12,000 x g for two minutes. The genomic region containing the mutated sites in the *sox32* locus was amplified using Q5 High-Fidelity DNA Polymerase (NEB, Cat#M0491) according to the manufacturer's instructions; 2µl of genomic DNA lysate was used in a 15µl reaction volume together with 200nM of forward and reverse primers (details at the end of chapter). The PCR cycling conditions were as follows: 2 minutes 98°C, 30 cycles (10 seconds 98°C, 30 seconds 58°C, 30 seconds 72°C) and 5 minutes 72°C. PCR products were purified by ethanol precipitation or by spin column technology using DNA Clean and Concentrator-25 (Zymo Research, Cat#D4033) according to the manufacturer's instructions. The presence of the mutation was verified by restriction fragment length polymorphism (RFLP), using BfaI restriction enzyme (NEB, Cat#R0568) as previously described (Alexander et al., 1999; Dickmeis et al., 2001). Fragments were resolved on a 2% agarose gel, using a 100bp DNA Ladder as a reference (NEB, Cat#N0467). The mutation introduces a restriction site for the BfaI enzyme, thus the mutant

allele yields two fragments of 129 and 123bp compared to the uncut wild type fragment of 252bp.

2.1.4 *sox32* mutant in-cross embryo genotyping

In order to genotype single embryos from *sox32*$^{+/-}$ in-crosses, individual embryos were collected and homogenised in 200µl TRIzol (Thermo Fisher Scientific, Cat#15596026), and genomic DNA and total RNA extracted according to the manufacturer's back extraction protocol. Briefly, for each developmental stage, single embryos were homogenised by pipetting using a P200 tip in a low retention microcentrifuge tube (Sigma-Aldrich, Cat#Z666548) and vortexed for 1-2 minutes. For phase separation, 40µl of 1-Bromo-3-Chloropropane (Sigma-Aldrich, Cat#B9673) was added to the homogenate, which was then shaken vigorously for 15 seconds, incubated for 10 minutes at room temperature and then spun for 10 minutes at 16,000 x g at 4°C. The aqueous phase containing total RNA was snap-frozen in liquid nitrogen and stored at -80°C, while the interphase-organic layer was processed to extract genomic DNA as follows: 250µl of back extraction buffer (Guanidine thiocyanate 4M, Tris 1M, Sodium citrate 50mM) was added, incubated for 10 minutes then spun for 30 minutes at 16,000 x g at room temperature. The aqueous phase was then transferred to a new tube and genomic DNA precipitated using 1 volume of isopropanol and 0.5µl of GlycoBlue coprecipitant (Thermo Fisher Scientific, Cat#AM9515) incubated for one hour at room temperature followed by centrifugation at 12,000 x g for 15 minutes at 4ºC. The DNA pellet was washed twice in freshly prepared ice cold 70% ethanol and resuspended in 10µl of nuclease free water. To help dissolve the DNA pellet, samples were incubated at 55ºC for 5 minutes. PCR was then performed as described above. PCR products were purified to remove remaining primers and dNTPs using Exonuclease I (NEB, Cat#M0293) and Shrimp Alkaline Phosphatase (NEB, Cat#M0371) (15 minutes 37°C, 15 minutes 80°C) and then sequenced using Sanger sequencing.

2.1.5 High-resolution melting (HRM)

Primers were designed using Primer3Plus (http://www.bioinformatics.nl/cgi-bin/primer3plus/primer3plus.cgi) according to Applied Biosystems recommendations: ~130bp amplicon length, 20bp primer length and primer melt temperature (Tm) around 60°C. Primers are reported at the end of the chapter. A single melting domain for the HRM fragment was checked using uMelt (Wittwer et al., 2011) and predicted melting curves of PCR products from the expected genotypes (wild type, heterozygous and homozygous) were obtained using uMelt HETS (https://www.dna.utah.edu/umelt/umelt.html). The PCR reactions comprised 5µl

of MeltDoctor HRM master mix (Thermo Fisher Scientific, Cat#4415440), 0.5µl each of forward and reverse primers (10µM), 2µl of genomic DNA and ddH_2O up to 10µl. The PCR was performed in a ViiA 7 system (Applied Biosystems) using 384 well plates. The two-step PCR reaction protocol used was 95°C for 10 minutes, then 40 cycles of 95°C for 15 seconds and 60°C for 60 seconds, followed by a melt curve/dissociation step between 60°C and 95°C. Curves were analysed using the ViiA 7 Software version 1.5.1.62.

2.2 Histological Techniques

2.2.1 Whole mount *in situ* hybridisation (WMISH)

Whole mount in situ hybridizations were performed according to standard protocols (Thisse and Thisse, 2008). Embryos were fixed in 4% [w/v] paraformaldehyde (Sigma-Aldrich, Cat#158127) in PBS (Phosphate buffered saline, Thermo Scientific Oxoid, Cat#10209252) overnight at 4°C. The embryos were then manually dechorionated, washed 3 times in PBS for 5 minutes and 2 times in methanol for 5 minutes. To ensure proper dehydration, fixed and dehydrated embryos were store at -20°C for at least 24 hrs. Digoxigenin-tagged antisense RNA probes were synthesised *in vitro* from either linearized cDNA clones or PCR product with the appropriate SP6 or T7 RNA polymerase (Promega C#P1085 and C#P2075) and digoxigenin-11-UTP (Roche, Cat#11277065910). Plasmid constructs and primer sequences for PCR for all probes are listed at the end of the chapter. After DNase treatment according to manufacturer's protocol (RQ1 DNase, Promega, Cat#M6101), probes were precipitated with LiCl overnight at -20°C. The RNA precipitate was rinsed twice with 70% ethanol then resuspended in RNase-free water. The RNA was then quantified using a DeNovix DS-11spectrophotometer (DeNovix) and diluted with hybridisation buffer (50% [v/v] formamide, 5x SSC (3M NaCl in 0.3M sodium citrate (pH 7.0)), 10mM EDTA, 1mg/ml torula RNA, 100µg/ml heparin and 0.1% [v/v] Tween-20) to a final concentration of 10ng/µl (10x stock) and stored at -80°C. For the WMISH protocol, embryos were transferred into baskets (maximum 50 per basket), which were kept in 100-place polypropylene storage box with fixed dividers (Starlab, Cat# E2100-5999) filled with the appropriate reagents. Baskets were built by replacing the round bottom of a 2ml microfuge tube with a Sefar Nitex mesh. The basket system was used to process several batches of embryos at the same time because all baskets can be transferred at the same time to submerge embryos into subsequent buffers of the protocol. Dehydrated embryos were rehydrated through a methanol series to PBT (1x PBS, 0.1% [v/v] Tween-20) in 3 steps of 5 minutes each (75% methanol/PBT, 50% methanol/PBT, 25% methanol/PBT). Embryos

younger than 24 hpf were then treated with 5µg/ml proteinase K (Thermo Fisher Scientific, Cat#AM2548) in PBT for 2 minutes, embryos 24 hpf and older were treated for 10 minutes. Embryos were then post-fixed in 4% [w/v] paraformaldehyde in PBT for 20 minutes, then washed 4 times in PBT with gentle rocking. Next, embryos were incubated in 650µl of hybridisation buffer for 4 hours in a hybridisation oven set to 60°C. After this pre-hybridisation step, the embryos were transferred into 500µl of 1ng/µl digoxigenin-labeled probe in hybridisation buffer (preheated to 60°C) and incubated overnight at 60°C. The following morning, the probe was removed and stored at -80°C for future re-use and the embryos were transferred into fresh pre-warmed hybridisation buffer and incubated at 60°C for 10 minutes. The following washes were then undertaken: 2 times in 2x SSC/0.1% [v/v] Tween-20 for 15 minutes at 60 °C, 2 times in 0.2x SSC/0.1% [v/v] Tween-20 for 30 minutes at 60°C and once in maleic acid buffer (1 x MAB, 0.1M maleic acid, 0.15M NaCl,) (pH 7.5)) for 15 minutes at room temperature. Next, the embryos were incubated in blocking solution (2% [w/v] Boehringer Mannheim blocking reagent (Roche, Cat#11096176001), 10% [v/v] lamb serum in 1x MAB) for 30 minutes at room temperature and then left overnight at 4°C in antibody solution (1:2000 anti-digoxigenin antibody coupled to alkaline phosphatase (Anti-Digoxigenin-AP Fab Fragments, Roche Cat#11093274910), 10% [v/v] lamb serum, 2% [w/v] Boehringer Mannheim blocking reagent (Roche, Cat#11096176001), 1x MAB). On the final day of the protocol, excess antibody was removed by extensively washing the embryos in MBT (1% [v/v] Tween-20 in MAB) at least five times over 3 hours before equilibrating the embryos in freshly prepared AP buffer (50mM $MgCl_2$, 100mM NaCl, 100mM Tris pH 9.5, 1% Tween-20). The colorimetric reaction was visualised by incubating the embryos in freshly prepared staining solution (AP buffer, 338µg/ml NBT (nitro-blue tetrazolium chloride), 175µg/ml BCIP (5-bromo-4-chloro-3'-indolyphosphate)) at room temperature and protected from light until the staining had sufficiently developed. The reaction was then stopped by 2 washes of 5 minutes in PBT. To stabilise the signal and preserve morphological features, the embryos were then washed twice in 70% methanol/PBT and cleared by replacing the methanol with BABB (1-part benzyl alcohol, Sigma-Aldrich, Cat# B1042), 2-parts benzyl benzoate (Sigma-Aldrich, Cat# B6630) also known as “Murray's reagent” and imaged as described later.

2.2.2 Anti-GFP immunostaining

Embryos were fixed and dehydrated as outlined above for WMISH. All subsequent protocol steps were performed at room temperature with gentle rocking unless otherwise stated.

Dehydrated embryos were transferred into 2ml round bottom microcentrifuge tubes and washed in ddH_2O for 5 minutes, permeabilised in ice cold acetone for 7 minutes, washed again in ddH_2O for 5 minutes then equilibrated in PBT. Next, embryos were blocked for 1 hour in PBS-SSDT (PBS, 2% goat serum, 0.1% [v/v] Triton X-100, 1% [v/v] DMSO, 1% [v/v] BSA (Bovine Serum Albumin)) before incubation in primary antibody (1:300 mouse anti-GFP, Invitrogen Cat#A11120) in PBS-SSDT for either 4 hours at room temperature or overnight at +4°C. Subsequently, embryos were rinsed multiple times in PBS-DT (PBS, 0.1% [v/v] Triton X-100, 1% [v/v] DMSO) for at least 3 hours, then incubated in secondary antibody solution (1:400 goat anti-mouse IgG HRP-conjugated, 0.4mg/ml, ThermoFisher Cat#31430) in PBS-SSDT for 4 hours. Embryos were then rinsed thoroughly in PBT before the HistoGreen colorimetric reaction was carried out according to the manufacturer's instructions (Linaris, Cat#E109). For imaging, embryos were passed through a glycerol series with 5 minutes equilibration steps (30% glycerol/PBT, 50% glycerol/PBT, 70% glycerol/H_2O) and imaged as described below.

2.2.3 Image acquisition and processing

Bright field images and WISH image were obtained either using a digital camera (UI-3080CP-C-HQ; IDS Imaging Development Systems) attached to a Leica MZ125 stereomicroscope or using a Leica DFC310 FX digital camera with a Leica M165 FC stereo microscope. Images were processed in FIJI/ImageJ (Schindelin et al., 2012) and brightness, sharpness, contrast and colour balance were applied uniformly to images using Adobe Photoshop. For quantification of WISH signals, negative images of each embryo were made. Each yolk was then individually outlined, and the "Histogram" information of each selection obtained including "Mean" and "Pixels" values. To give a measure of absolute intensity, the mean value was multiplied by the pixel value. To acquire images of *tg(sox17:GFP)* embryos, an ExiAqua (Q-imaging) camera on a Nikon SMZ1500 microscope was used. Exposure, gain, brightness and colour setting were adjusted using wild type (non-fluorescent) fish.

2.3 Gene Expression Analysis – *sox17:GFP* transgenic line

2.3.1 RNA extraction and cDNA synthesis from single embryos

For each developmental stage, individual embryos were homogenised using a P200 filtered tip in 200µl TRIzol (Thermo Fisher Scientific, Cat#15596018), left at room temperature for

10 minutes, then centrifuged for 15 minutes at 4°C at full speed. 20µl BCP (1-Bromo-3-chloropropane, Sigma Aldrich, Cat#B9673) was added to the homogenate which was then shaken vigorously for 15 seconds, then kept on ice for 5 minutes. The partly separated mixture was transferred to a clean 1.5ml low retention microcentrifuge tube and centrifuged for 15 minutes at 12,000 × g. The aqueous phase was then mixed with 1 volume of absolute ethanol and the RNA purified by column precipitation using the Zymo RNA Clean and Concentrator-5 kit (Zymo Research, Cat#R1013) according to the manufacturer's instructions, with in-column 3U DNAse treatment (Thermo Fisher Scientific, Cat#AM2238). At the end of the protocol, the RNA was eluted in 15µl nuclease-free water and stored at -80°C until needed. The RNA concentration was measured on the Qubit fluorometer (Life Technologies) with the Qubit RNA HS Assay Kit (Invitrogen, Cat#Q32852) and the purity with the DeNovix DS-11spectrophotometer (DeNovix). For the 'non-leaky' and 'leaky' experiment, RNA was extracted from 3 batches of 3 individual embryos per condition; for the wild type control experiment RNA was extracted from 1 batch of 3 individual embryos. The overall experimental approach is shown in Figure 2.1

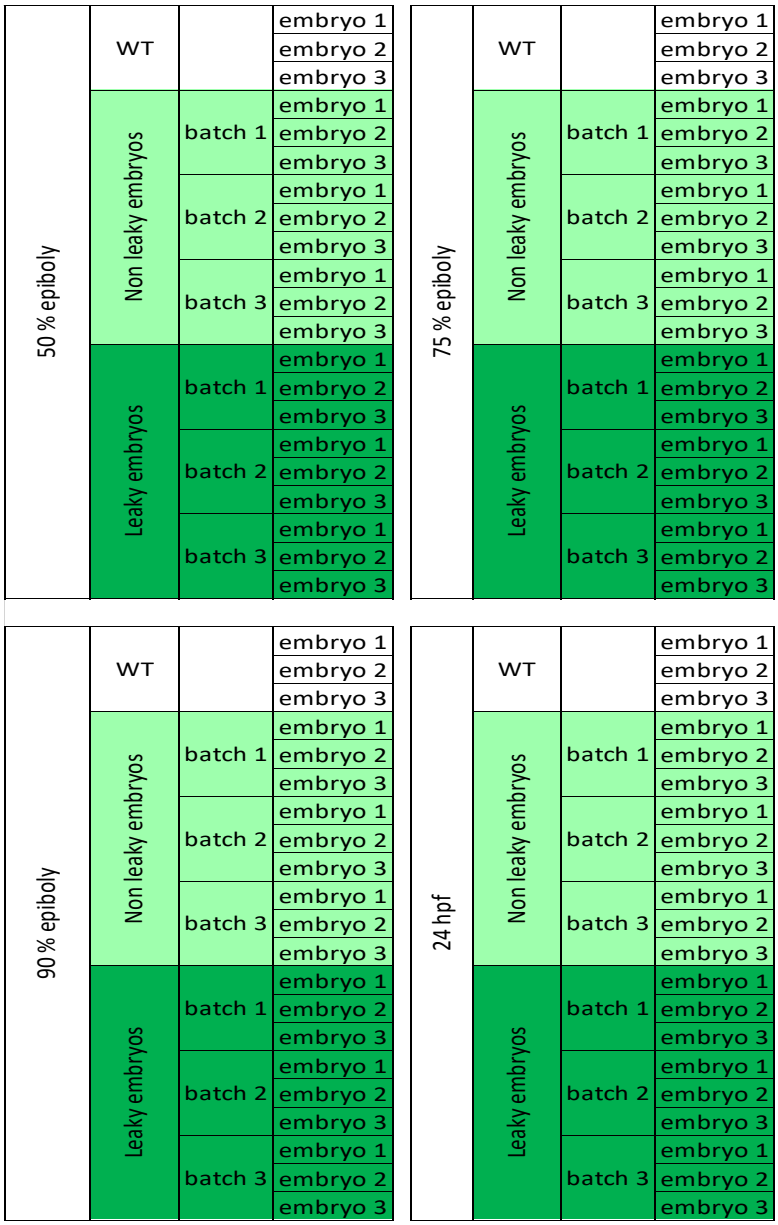


Figure 2.1 Experimental approach to characterise the gene expression in 'leaky' and 'non leaky' embryos at 4 different developmental stages.

2.3.2 Quantification of transcription (RT-qPCR)

Approximately 150ng of total RNA was reverse transcribed (RT) using the high-capacity cDNA reverse transcription kit (Thermo Fisher Scientific, Cat#4368814), in a 20µl reaction under the following conditions: 15 minutes at 25°C, 120 minutes at 37°C and 15 minutes at 85°C. The resultant cDNA was then diluted with molecular grade H_2O to a concentration of 0.4ng/µl for qPCR. The qPCR reactions (1.2ng cDNA per reaction) were run in technical duplicates using Luna Universal qPCR Master Mix reagents (NEB, Cat#M3003) and the ABI PRISM ViiA7 sequence detection system (Applied Biosystems), cycling 40 times between

95°C (15 seconds) and 60 °C (60 seconds). This was followed by a melt curve step with a temperature ranging from 60 to 95°C to confirm the presence of a single specific amplicon. The standard curve for each primer pair was calculated from 5 crossing points (C_t-values), generated from the amplification of a 1:5 serial dilution of wild type cDNA (slope values between -2.95 and -3.75). Relative gene expression levels were then determined using the absolute quantification methodology and normalised to the reference (housekeeping) gene elongation factor-1 alpha (*ef-1α*), by the generation of a calibration curve as previously described (Larionov et al., 2005). *ef-1α* was considered a suitable reference gene because mRNA expression was stable across the range of development times (Tang et al., 2007). The primer sequences used are listed at the end of the chapter. Student's two-tailed t-tests were performed for pairwise comparisons to determine any statistically significant differences between groups.

2.3.3 Fluorescence-activated cell sorting (FACS)

Embryos from homozygous *tg*(*sox17*:*GFP*) in-crosses were collected 20 minutes after fertilization. Selection of mating pairs was random from a pool of 21 males and 18 females. Embryos were dechorionated by incubation in 1mg/ml pronase (protease from Streptomyces griseus, Sigma-Aldrich Cat#11459643001) in a glass beaker for 5–6 minutes until chorions began to crumble, then rinsed with several washes of E3 medium (5mM NaCl, 0.17mM KCl, 0.33mM $CaCl_2$, 0.33mM $MgSO_4$, dissolved in water). Embryos were then kept in the 33°C incubator in plastic petri dishes coated with 1% agarose until they reached the desired developmental stage. Using a Nikon SMZ1500 dissecting microscope with a fluorescent filter set, dishes were visually inspected and 'non leaky' and 'leaky' carriers of the transgene were separated. Embryos were then placed into separate 1.5ml low retention microcentrifuge tubes, as much E3 medium as possible was removed and 200µl of filtered ice cold 10% BSA (Bovine Serum Albumin, Sigma-Aldrich Cat#A2058) in HBSS solution without calcium or magnesium (Thermo Fisher Scientific, Cat#14170088) was added. Samples were processed in parallel and care was taken to ensure that no cells were lost during the following steps. Dissociation was facilitated by gentle pipetting using low retention P200 pipette tips with a filter barrier to avoid contamination. For all samples, the vital dye DAPI (Insight Biotechnology, Cat#AR1176) was added at 50µl per 150µl cell suspension. The dissociated cells were then pelleted using a table-top centrifuge at 300 x g for 4 minutes at 4°C. The supernatant was removed and the pellet thoroughly resuspended in 1ml of filtered ice cold 10% BSA in HBSS. The samples were kept

on ice prior to flow cytometry. Each cell suspension sample was then filtered into a Falcon tube using a cell strainer cap (a 35µm nylon mesh tube cap on a falcon collection tube (Thermo Fisher Scientific, Cat#10585801)) prior to sorting, to remove any unwanted clumps of cells. In order to obtain as many cells as possible, the tube used for dissociation was then rinsed with 0.5ml of filtered 10% BSA in HBSS. The minimum sorting volume used was 1.5ml. GFP positive and negative cells were sorted directly into ice cold PBS using the BD FACS Aria II (BD Bioscience) equipped with a 100µm nozzle at King's College London Flow Cytometry & Cell Sorting Facility. For the FACS experiment, I used the following controls: (1) negative, unstained sample (WT embryos, no DAPI added); (2) DAPI compensation negative control (WT embryos, with DAPI); (3) negative unstained sample (*tg*(*sox17*:*GFP*) embryos, no DAPI added) and (4) positive fluorophore control (*tg*(*sox17*:*GFP*) embryos with DAPI). Prior to running the experimental samples, we optimised and adjusted the gating of the FAC sorter using both the WT (no DAPI) and the DAPI compensation control. This allowed me to determine the level of background fluorescence and autofluorescence and to set the voltages and negative gates appropriately.

2.3.4 RNA extraction from sorted cells

All equipment used for RNA isolation was cleaned with RNAse Away (Thermo Fisher Scientific, Cat#10666421) and filter tips were used to avoid contamination. The sorted cells were pelleted using a benchtop centrifuge at 300 x g for 10 minutes at 4°C, the supernatant removed and a variable volume of TRIzol (Thermo Fisher Scientific, Cat#15596018), depending on the number of cells collected, was added. The ratio of TRIzol to sample was optimised starting from the manufacturer's recommendations: 50,000 cells – 200µl, 100,000 cells – 300µl, 150,000 cells – 300µl, 200,000 cells – 400µl and 250,000 cells – 400µl. Cells were lysed by gently pipetting the homogenate several times and then incubated on ice for 10 minutes before being snap-frozen in liquid nitrogen and stored at -80°C. RNA was extracted by adding 10% BCP (1-Bromo-3-chloropropane, Sigma Aldrich, Cat# B9673) (e.g. 20µl BCP to 200µl TRIzol) to the cell lysate, shaking the tubes vigorously for 15 seconds and then spinning for 15 minutes at 16,000 x g at 4°C. The RNA in the aqueous phase was then precipitated out using one volume of absolute ethanol and cleaned using the RNA Clean and Concentrator-5 (Zymo Research, Cat#R1013) according to the manufacturer's instructions, with in-column 5 U Turbo DNase (Thermo Fisher Scientific, Cat#AM2238) treatment for 10 minutes. The RNA was stored at -80°C until needed. RNA quantity was determined by fluorometry using Qubit

RNA HS reagents (Invitrogen, Cat#Q32852) and genomic contamination was assessed by PCR using GoTaq G2 Flexi DNA Polymerase (Promega, Cat#M7801) and 0.25 µM of primers (Actb2-RT-F:5-GCCCCTAGCACAATGAAGAT-3, Actb2-RT-R:5-GTTTGAGTCGGCGTGAAGT-3) for the *18s* gene spanning exons 2 and 3. PCR cycling conditions were as follows: 95°C for 5 minutes, then 35 cycles of 30 seconds at 95°C, 30 seconds at 60°C then 30 seconds at 72°C, ending with 72°C for 5 minutes. PCR products were analysed by size using agarose gel electrophoresis; pure RNA showed no band, DNA-contaminated RNA produced a band at 285bp and amplification from the cDNA sample yielded a 200bp band. RNA quality and integrity were analysed using the RNA nano chip (Agilent, Cat#5067-1511) on an Agilent 2100 Bioanalyzer according to manufacturer's instructions.

2.3.5 RNA extraction for RT-qPCR and sequencing

Embryos from *tg*(*sox17:GFP*) in-crosses were manually sorted to separate non-leaky and leaky carriers, and prepared for FACS as describe above. GFP high and GFP low cells were sorted according to gating conditions described further in the Results section. A typical sort generated 50,000 cells for RT-qPCR analysis and 50,000 cells for RNA-seq library preparation. Reverse transcription was carried out using the high-capacity cDNA reverse transcription kit (Thermo Fisher Scientific, Cat#4368814) according to the manufacturer's instructions. qPCR TaqMan probes were designed by Applied Biosystems to precisely quantify transcript abundance of known endodermal, mesodermal and ectodermal genes (see end of the chapter for primers description). All target gene probes were labelled with FAM dye whereas the reference gene (18S) was labelled with VIC dye, to enable multiplexing in the same reaction solution. To ensure optimal multiplexing, the melting temperatures of the target primers were similar to those of the 18S probes. To verify the assay's capability for simultaneous detection of different markers in same sample, we compared Ct values of multiplex vs singleplex in different combinations: whole embryo cDNA, low GFP cDNA and high GFP cDNA. The RT-qPCR reaction consisted of 1ng of cDNA and 5µl of TaqMan Gene Expression Master Mix (Applied Biosystems, Cat#4369016) in a final volume of 10µl. Amplification conditions were 2 minutes at 50°C followed by 10 minutes at 95°C, and then 40 cycles of 15 seconds at 95°C and 1 minute at 60°C. Gene expression levels were calculated relative to the reference gene 18S, using the $2^{-\Delta\Delta Ct}$ method (Livak and Schmittgen, 2001). The mean and standard error of

the mean (SEM) were plotted for each condition. 3 technical replicates in addition to 4 biological replicates were used.

Poly(A)+ RNA-Seq libraries were made from ~100 ng total RNA extracted from non-leaky embryos (5 biological replicates) and leaky embryos (1 biological replicate) and using the NEBNext Poly(A) mRNA magnetic isolation module and the NEBNext Ultra II Directional RNA Library Prep Kit for Illumina (NEB, Cat#E7760), both according to the manufacturer's instructions. Libraries were sequenced on an Illumina HiSeq 4000 platform at BGI (Honk Kong) to produce paired-end reads of 50 bases.

2.4 Gene Expression Analyses – *sox32* mutant embryos

2.4.1 RNA extraction from *sox32*$^{-/-}$ embryos

For deep sequencing, individual embryos from separate *sox32*$^{+/-}$ in-crosses were collected at developmental stages 5.25 hpf (50% epiboly) and 9.00 hpf (90% epiboly). Each embryo was homogenised in 200µl TRIzol by pipetting (to break the chorion) and vortexing. For phase separation, 20µl of BCP was added to the homogenate, which was shaken vigorously for 15 seconds before spinning for 5 minutes at 16,000 x g at 4°C. The aqueous phase containing total RNA was snap-frozen in liquid nitrogen and stored at -80°C, while the interphase-organic layer was processed to extract genomic DNA as described above for genotyping. The aqueous phases from 3 sibling embryos of common genotype (*sox32*$^{+/+}$ and *sox32*$^{-/-}$) were then combined, before being precipitated with one volume of absolute ethanol and cleaned using the RNA Clean and Concentrator 25 (Zymo Research, Cat#R1017) with in-column 5U Turbo DNase (Thermo Fisher Scientific, Cat#AM2238) treatment for 10 minutes as above. The quality and concentration of extracted RNA was estimated using the RNA assay program on a Qubit fluorometer (Life Technologies), while the integrity of the RNA was investigated by Bioanalyzer (Agilent Technologies) using the RNA nano kit (Agilent Technologies). The cut-off for the RIN value was set at 7.

2.4.2 Poly(A)+ RNA-Seq library preparation and sequencing

Triplicate biological libraries for both stages described above were prepared using NEBNext Poly(A) mRNA magnetic isolation module in combination with NEBNext Ultra II Directional RNA Library Prep Kit for Illumina (NEB, Cat#E7760), both as per the manufacturer's

instructions. To generate complex cDNA libraries, we used the minimum recommended PCR cycles based on total RNA input amount (500 ng). The Qubit fluorometer (Invitrogen) and Agilent 2,100 Bioanalyzer (Agilent Technologies) were used to ensure the inserts were both the appropriate size, and to determine the concentration prior to sequencing, respectively. Transcriptome libraries were then sequenced using the HiSeq4000 Sequencing System (Illumina) at BGI (Hong Kong). Libraries were multiplexed and 12 samples per 50bp paired-end were sequenced, generating approximately 20 million reads per sample.

2.4.3 Target validation by RT-qPCR

For RT-qPCR, three *sox32*$^{+/+}$ and *sox32*$^{-/-}$ for each developmental stage were collected from separate fertilisations and processed as described above. 100ng of total RNA were reverse transcribed to cDNA using the high-capacity cDNA reverse transcription kit (Thermo Fisher Scientific, Cat#4368814), according to the manufacturer's instructions. Primers used to detect novel genes are listed at the end of the chapter. RT-qPCR was then performed on 2.5ng of cDNA using Luna Universal qPCR Master Mix reagents (NEB, Cat#M3003) on an ABI PRISM ViiA7 machine (Applied Biosystems). Triplicate cycle thresholds were normalised against the expression of the reference gene ef-1α and relative transcript levels calculated using the $2^{-\Delta\Delta Ct}$ method (Livak and Schmittgen, 2001). Statistical significance (p-value) was calculated using students' t-tests comparing transcript levels in mutant embryos to control wild type siblings.

2.5 Gene Expression Analysis – *mixl1* mutant embryos

2.5.1 RNA extraction from *mixl1*$^{-/-}$ embryos

Homogenised *mixl1*$^{-/-}$ embryos (3 biological replicates of 50 embryos each) in TRIzol were a kind gift from Prof Dirk Meyer and Dr Patrick Fischer at the University of Innsbruck. In addition, further RNA extraction from *mixl*$^{-/-}$ embryos was undertaken in London (see Results section for rationale). To do so, 50 embryos from *mixl1*$^{-/-}$ in-crosses were homogenised using a Tissue Raptor (Qiagen) in 1000µl of TRIzol. The homogenate was then mixed with 100µl of BCP and centrifuged for 15 minutes at 4°C at full speed. The upper phase was transferred to a clean microcentrifuge tube. RNA was precipitated using the RNA Zymo Clean and Concentrator-25 kit with in column DNAse treatment as previously described. The quality of the extracted RNA was estimated by absorbance ratios (A_{260}/A_{280} 1.8–2.0 and A_{260}/A_{230}>1.7)

using a DeNovix DS-11 spectrophotometer (DeNovix) and the concentration was determined using the RNA assay program on a Qubit fluorometer (Life Technologies). For RT-qPCR validation, approximately 100ng of total RNA was transcribed into first strand cDNA using the Thermo Fisher high-capacity cDNA reverse transcription kit according to the manufacturer's instructions. qPCR reactions were run on ABI PRISM ViiA7 machine (Applied Biosystems) using Luna Universal qPCR Master Mix reagents (NEB, Cat#M3003), see primer sequences at the end of the chapter and Ct values and the relative level of gene expression between mutant and wild type embryos was analysed using the $2^{-\Delta\Delta Ct}$ method as described previously.

2.5.2 Ribosomal RNA depletion, RNA-Seq library preparation and sequencing

Genome-wide transcriptome libraries for each condition (*mixl1*$^{-/-}$ and wild type) were generated from both samples received from Innsbruck and sample prepared in London. This therefore totalled 6 wild type control and 6 *mixl1*$^{-/-}$ libraries. Total RNA (4µg input) representing 50 *mixl1*$^{-/-}$ embryos or 50 wild type *Tuebingen* embryos (a kind gift from the Francis Crick Institute Aquatics) was processed using the Ribo-Zero rRNA Removal Kit (H/M/R) (Illumina, Cat# MRZH116) to deplete ribosomal RNA followed by NEBNext Ultra II Directional RNA Library Prep Kit for Illumina (NEB, Cat#E7760), both according to the manufacturer's instructions. Library concentration was determined by fluorometry using Qubit dsDNA HS reagents (Thermo Fisher Scientific) and Agilent 2,100 Bioanalyzer (Agilent Technologies) was used to determine quality prior to sequencing as previous described. Size selection of the library using SPRIselect Beads (Beckman Coulter, Cat#B23317) was performed if the library displayed a 127 bp adaptor/dimer peak on the Bioanalyzer trace. Libraries were sequenced with the Illumina HiSeq4000 (2 × 50 bp) at BGI (Hong Kong).

2.6 ChIP-exo libraries

2.6.1 Western blot analyses and *in vitro* protein production

40 to 50 embryos were dechorionated, snap frozen and stored at -80°C until processing. Pooled embryos were then homogenized in lysis buffer (20mM Tris HCl pH 8, 2mM EDTA pH 8, 0.5% NP-40, 25mM β-glycerophosphate, 100mM NaF, 20 nM Calyculin A, 100mM sodium pyrophosphate and protease inhibitors). The lysates were loaded onto a standard SDS/PAGE 10% gel. After electrophoresis, proteins were transferred to PDVF membrane (GE

Healthcare, Cat#10600069) and immunoblotted using standard protocols. The following primary antibodies were used: anti-Sox32 (rabbit polyclonal, raised against the C-terminal region of zebrafish Sox32 protein, GenBank accession NP_571926.1, Anaspec Cat# AS-55856, lot JI2801); anti-Sox17 (rabbit polyclonal, raised against the intermediate region of zebrafish Sox17 protein, GenBank accession NP_571362.2, Anaspec Cat# AS-55856, lot JI1508); anti-Mixl1 (rabbit polyclonal, raised against the intermediate region of zebrafish Mixl1 protein, GenBank accession NP_571015.2, Anaspec Cat# AS-55613, lot JJ1505 and JJ1502 (Nelson et al., 2017)), with concentration as reported in each Figure in Chapter 3. Blots were then washed extensively in TBST, incubated for 1 hour in goat anti-rabbit HRP-conjugated antibodies (GE Healthcare) and washed extensively in TBST for 4 hours. Peroxidase activity was detected with SuperSignal West Pico Rabbit IgG detection kit (Thermo Fisher Scientific, Cat#3408) according to manufacturer's instructions and captured with a ChemiDoc XRS+ imager (BioRad).

The full-length coding region of the zebrafish *sox17*, *mixl1* and *sebox* genes was amplified from cDNA obtained from 5.25 hpf embryos. The coding sequences were first inserted into pBluescript KS (-) vector (available in the lab) then sub-cloned unidirectionally into BamHI/NotI sites of the pCS2+ expression vector. The pCS2+ vectors were then linearised with NotI and transcribed *in vitro* in order to synthesise the respective mRNA. The *sox32* pCS2+ plasmid was previously described (Nelson et al., 2017). Plasmids containing the full cds sequence of *sox7*, *sox10*, *sox18* were kind gifts from (Dutton et al., 2001; Swift et al., 2014). The primers used for amplification are listed at the end of the chapter. Sox32, Sox17, Sox7, Sox10, Sox18, Mixl1 and Sebox proteins were synthesised as follows: 0.5 µg of the respective mRNA was incubated with rabbit reticulocyte lysate (Promega, Cat# L4960) for 60 minutes at 37°C. For immunoprecipitation, 10µl of Dynal magnetic beads (Invitrogen, USA) were coated with 5µg of the appropriate antibody described previously, The beads were then incubated with 3µl of the reticulocyte lysate and 200µl IP buffer (16.7 mM Tris-HCl pH 7.5, 167 mM NaCl, 1.2 mM EDTA, 0.01% SDS) and mixed gently overnight at 4°C. Beads were then washed four times with ice-cold RIPA buffer (50mM Hepes (pH 7.6), 1 mM EDTA, 0.7% DOC, 1% IGEPAL, 0.5 M LiCl), followed by elution in TE buffer pH 8.0 on a shaker for 60 minutes at 60°C. Beads were recovered by centrifugation at 800 × g for 3 minutes. 20µl of 2× SDS-PAGE sample buffer was then added to the supernatant and heated to 95°C for 10 minutes. Samples were size-separated by electrophoresis in SDS-containing (10%) polyacrylamide gels and transferred to PDVF membranes, then blocked at 4°C for 30 minutes

with PBT containing 5% skim milk powder. Membranes were then incubated at 4°C overnight with the primary antibodies (as above) at the following concentrations: Sox32 (1:1,500), Sox17 (1:1,500), Mixl1 (1:1,000). Membrane were washed 8 times (10 minutes each) with PBST. For detection, blots were incubated with anti-rabbit IgG-HRP, for 1 hr at room temperature. After washing 5 times at room temperature with PBST (10 minutes each), immune-reactive bands were detected on autoradiography film.

2.6.2 HEK293 transfection assays

sox32 cDNA was amplified from the *sox32* pCS2+ plasmid (see above) and subcloned using Gilson assembly into a pEGFP-C1 vector (Clontech, kindly provided by Andrea Ghisleni), to produce an N-terminal tagged *egfp-sox32* fusion construct. The primers used are listed at the end of the chapter. Transactivation assays were performed by co-transfecting HEK293 cells with the *egfp-sox32* expression vector and cell lysates were subject to western blot analyses as described in the previous section.

2.6.3 Chromatin shearing

Sonication conditions need to be optimised for different experiments and for each different sample type. Therefore, to determine the optimum sonication time for this assay, a time course experiment was performed on 1,500 5.25-9.00 hpf zebrafish embryos. Briefly, embryos were dechorionated with pronase (Sigma, Cat#11459643001) and fixed in 1.85% formaldehyde in E3 medium for 25 minutes at room temperature. A 1/20th volume of glycine (2.5M) was added to quench the formaldehyde for 5 minutes and the embryos were washed in ice-cold PBS 4 times. Fixed embryos were homogenised in lysis buffer (10mM Tris-HCl pH 7.5,10mM NaCl, 0.5% NP-40) and incubated on ice for 20 minutes. Nuclei were collected by centrifugation and resuspended in nuclei lysis buffer (50mM Tris-HCl pH 7.5, 10mM EDTA, 1% SDS) for 10 minutes on a shaker, before diluting with 2 parts IP buffer (16.7mM Tris-HCl pH7.5, 167mM NaCl, 1.2mM EDTA, 0.01% SDS) to one part nuclei lysis buffer. Protease inhibitors (cOmplete Mini, EDTA-free Protease Inhibitor Cocktail, Roche, Cat# 04693159001) were add to all buffers before use. 50µl of nuclear lysate was stored without sonication; the rest was sonicated over a time-course to identify optimal conditions to give fragments of 200-700bp. Samples were processed for 1, 2, 3, 4, 5, 6, 7, 8 and 9 minutes. 20µl samples were removed each sonication round and DNA isolated. The fragment size decreased over the time course, with the optimal fragment size observed at 8 minutes (cycles of 30 seconds ON/30 seconds

OFF on high (Bioruptor Diagenode, Be). 100µl of Tris (25 mM, pH 9.8) with EDTA (1mM), RNAse A 0.2µg/µl (Thermo Fisher Scientific, Cat#EN0531) and 20µg of Proteinase K (Thermo Fisher Scientific, Cat#25530049) was added to the samples and which were then incubated at 55°C for 15 minutes. DNA was eluted from the beads at 100°C for 15 minutes, then centrifuged at 4°C for 3 minutes at 800 x g. Prior to analysis on a 1% agarose gel, chromatin fragments were cleaned up using Zymo ChIP DNA Clean & Concentrator (Zymo, Cat# D5201) or purified by phenol:chloroform:isoamyl alcohol extraction followed by ethanol precipitation.

2.6.4 Sample preparation and ChIP-exo sequencing

Two independent ChIPs were carried out for each antibody (Sox32, Sox17, Mixl1 – 25 µg per ChIP). ChIP-exo assays were each performed using 9000 embryos respectively at 50% epiboly (5.25 hpf) and 90% epiboly (9.00 hpf); embryos from several different crosses on several different days were pooled for large scale chromatin immunoprecipitation. DNA extraction for the IP was in accordance with the protocol described in (Morley et al., 2009). Briefly, embryos were dechorionated with pronase, rinsed with E3 medium and then crosslinked for 25 minutes with 1.85% formaldehyde at room temperature. A 1/20th volume of 2.5M glycine was added to quench the formaldehyde and the embryos were washed in ice-cold PBS before freezing at -80°C. Embryos were then lysed, and the nuclei were isolated and disrupted to release chromatin. Chromatin sonication conditions were 8 cycles of 30 seconds ON/30seconds OFF on high (Bioruptor Diagenode), as detailed above. 20µl of nuclear lysate from each sample was stored at -20°C as input control. The rest was incubated overnight with 100µl Dynal magnetic beads (Thermo Fisher Scientific, Cat#10007D) prebound with 50µg of the respective antibody. At the end of the ChIP protocol, when the chromatin/antibody complex was still linked to the magnetic beads, a High-resolution library preparation kit (Diagenode, Cat# C05010023) was used to prepare the ChIP-exo DNA libraries. Briefly, a P7-adaptor was ligated to the immunoprecipitated DNA. A lambda exonuclease (10U) then digested the DNA fragments starting from the exposed 5' end and stopping at the protein-DNA boundary. This eliminated the P7 adaptor sequence at the 5' end of each strand. Proteinase K (100µg) was added and the DNA was incubated for 4 hours at 65°C to reverse the cross-links. The eluted single-stranded DNA was then made double-stranded by P7 PCR primer extension (at the 3' end) prior to ligation of the second adaptor (P5). The resulting DNA was enriched by 13 cycles of PCR using NEBNext High-Fidelity PCR Master Mix (NEB,

Cat#M0541), according to the manufacturer's instructions. Quality and concentration of the resulting library was then analysed by Qubit High Sensitivity DNA Analysis Kits (Agilent Technologies) and KAPA Library Quantification Kits (Kapa Biosystems, Cat# KK4824) respectively. Cluster generation was performed using the Illumina Cluster Reagents preparation, and the library was sequenced on the HiSeq4000 platform (Illumina) with a rapid run to generate 50bp paired end reads at BGI (Honk Kong).

2.6.5 ChIP-qPCR

ChIP-qPCR assays were performed using 5.25 hpf and 9.00 hpf embryos (500-800 embryos per biological replicate). Chromatin was prepared and sheared by sonication for 7 cycles (30sec ON/30sec OFF) to a range of 0.4 to 0.7 kb as described earlier. The equivalent of 60 µg of chromatin was immunoprecipitated with 3µl of anti-Sox32 or anti-Mixl1 (as previously described) or mock antibody (IgG). 10% of the chromatin was retained as input before IP and used for qPCR standard curves. I used the *rhod* promoter as a negative control region (Morley et al., 2009). All the primers used in these experiments are reported at the end of Chapter.

2.7 Bioinformatics analysis

2.7.1 ChIP-seq and ChIP exo pipeline

Raw 50 bp paired-end reads generated from the Illumina HiSeq4000 were first assessed for sequence quality using the FastQC program (Andrews, 2010). Trimgalore software was used to remove adapter sequences and trim bases of low quality (Krueger, 2012). Trimmed raw files were inspected again using FastQC and then the reads were aligned to the zebrafish genome (build Zv10) using Bowtie (Langmead, 2010), allowing 2 mismatches and keeping only unique mapped reads. Unmapped reads were truncated by 6 bp from the 5' end and re-aligned. SAM files were converted to BAM files using Samtools (Li et al., 2009), and PCR duplicates were removed with Picard tools ((http://broadinstitute.github.io/picard/). IDR (Li et al., 2011) was run to calculate common peaks in files of biological replicates and MACS2 (Feng et al., 2011) and GEM (Guo et al., 2012) were used to calculate the fold enrichment (peak calling). Common peaks were identified using findOverlaps' command from the Bioconductor package 'GenomicRanges' (Lawrence et al., 2013) and the BEDTools was used to create bigwig file (Quinlan and Hall, 2010). HOMER (Heinz et al., 2010) was used to find the closest gene to each peak and *de novo* motif search and DAVID (Huang da et al., 2009)

was used for the enrichment analysis. Coverage of uniquely mapped reads was normalized by number of mapped reads and converted to BigWig format using deepTools. Tracks were visualized in the IGV genome browser (Robinson et al., 2011).

2.7.2 RNA-seq pipeline

The quality control and adaptor trimming of raw FastQ files was performed using FASTQ (Andrews, 2010) and Trimgalore (Krueger, 2012). Trimmed raw files were re-inspected using FastQC and then were mapped with STAR (Dobin et al., 2013) to the same genome version used for ChIP-exo. The quantMode GeneCounts function was used to calculate per gene count. The raw read counts from each experiment were imported into R and differentially expressed genes between selected conditions were calculated with DESeq2 (Love et al., 2014). Differentially expressed genes were defined as having a minimal 1-fold difference compared to controls (absolute log2FoldChange > 1) and a FDR p ≤0.05 or p ≤0.01. The gene set enrichment analysis was carried out independently of both up-regulated and down-regulated genes using online g:Profiler (Peterson et al., 2016), PANTHER (Muruganujan et al., 2018) and ZEOGS (Prykhozhij et al., 2013). Normalized coverage BigWig tracks for RNA-seq data were generated from the resulting BAM files using bamCoverage from the deepTools package (Ramirez et al., 2014).

2.8 Statistics

Quantitative data are expressed as mean ± standard error. Numbers of biological replicates are reported in each figure. The statistical significance was determined using Prism software (GraphPad, version 8). To compare two groups, a Student's test (two-tailed) was applied whereas one-way ANOVA was performed for multiple groups. Significance levels are indicated by *($p < 0.05$), **($p< 0.01$), ***($p< 0.001$) and ****($p< 0.0001$).

2.9 Gene regulatory network

The topology of the endodermal gene regulatory network (GRN) model was visualized using the computational and graphical platform BioTapestry (Longabaugh, 2012). Regulatory relationships amongst genes are based on three sets of evidence for this work: ZFIN database, RNA-seq time series, ChIP-seq experiments. Spatial and temporal expression of genes associated with the term "endoderm" and "endodermal-like" were downloaded from ZFIN

(https://zfin.org/downloads, (Ruzicka et al., 2015)) and changes in RNA expression of these genes were evaluated in the context of gain- and/or loss-of-function experiments (injection of mRNA encoding a specific transcription factor, or a translation blocking antisense morpholino oligonucleotide, respectively) to determine direct connections between the above "input genes" and their downstream target genes. RNA-seq time series data (White et al., 2017) were used to highlight the dynamic changes in gene expression during zebrafish development and evidence of direct physical interaction between a transcription factor and its target regulatory region was evaluated from ChIP-seq experiments. Information from multiple manuscripts published in the last decade combined with the analysis of time-series gene expression data, published ChIP-seq data and my data were then assembled together to update the network presented by Chan et al. in 2009.

Chapter 3 – Characterising endodermal protein–DNA binding events during zebrafish gastrulation

Chapter 3 highlights:

- First time application of ChIP-exo in zebrafish in an attempt to determine, with high resolution, the TF-bound regions in the zebrafish genome for Sox32, Sox17 and Mixl1.
- Viewpoint of the importance of evolutionary analysis in evaluating the specificity of antibodies against TFs to enrich for specific chromatin fragments.

3.1 Introduction

At the heart of endoderm specification is a gene regulatory network where conserved TFs play an important role. Graded Nodal and Fgf signalling controls the expression of these TFs that act as intermediaries in ultimately specifying the direction of the mesendodermal precursors (Kiecker et al., 2016; Poulain et al., 2006; Shen, 2007). Changes in gene expression and cell fate are established by selective repressive and inductive interactions between pairs of TFs, in particular Mix homeobox TFs (Mixl1 and Sebox) have been implicated in mesendoderm development (Kikuchi et al., 2000; Pereira et al., 2012; Poulain and Lepage, 2002) while establishment of *sox32* and *sox17* expression is needed for acquiring endodermal cells identity (Alexander et al., 1999; Dickmeis et al., 2001; Kikuchi et al., 2001). This leads to a distinct transcriptional code that defines endodermal from mesodermal progenitor domains as explained in Chapter 1, with the cells closest to the margin at the beginning of gastrulation becoming endodermal, whereas mesodermal identity is determined further away from the margin. While some players in this endodermal specification pathway are already known, other components and defining mechanisms continue to be studied, and there are many gaps in our knowledge. My goal was to identify the components and connections in this network that are downstream of Sox32, Sox17 and Mixl1 starting by identifying where in the genome these proteins bind *in vivo,* during gastrulation.

The zebrafish is a powerful model in which to carry out multidimensional analyses and it is relatively straightforward to manipulate gene function on a high throughput scale during

development, tissue formation and/or organ differentiation (Driever et al., 1994; Koster and Sassen, 2015). However, where in other vertebrates certain genes have a unique and well-characterised function, due to genome duplication, zebrafish exbibits more corresponding orthologues (Howe et al., 2013; Postlethwait et al., 2000). This opens up the possibility (and therefore further complication) of the genes either maintaining partially redundant roles or potential sub-regionalization of the function (Kassahn et al., 2009; Klüver et al., 2005). In one respect, the duplication of genes makes it more complicated to both create and examine mutant phenotypes where any effects may be masked by compensation. One way around this is to generate double knock-out or double morpholino-based knock-downs; although this can introduce stress and increase the risk of false positives and off-target effects (Gentsch et al., 2018; Rossi et al., 2015; Stainier et al., 2017). On the other hand, gene duplication leading to gain of new regulatory roles that can act at different stages of development or in multiple tissues offers us great insights into, and opportunities to better understand, the regulatory compensatory and cooperation mechanisms that make development so robust and reproducible (Garcia-Fernàndez et al., 2009; Lan and Pritchard, 2016; Wagner, 2008).

In some circumstances, various studies have shown the power of sub-functionalization of two orthologues resulting from gene duplication (He and Zhang, 2005; Winkler et al., 2003; Zecchin et al., 2007), where each gene has a different role due to regulation of different enhancers, leading to their expression in distinct cells populations. For examples, in late endodermal patterning, *nkx2.2a*, *jagged1b* and *pax6b* are required for the appropriate specification of the pancreas, whereas their duplicated counterparts (*nkx2.2b*, *pax6a* and *jagged1a*) are not implicated in pancreas specification (Kleinjan et al., 2008; Pauls et al., 2007).

Sox17 is a master regulator of endoderm specification in *Xenopus* (Sinner et al., 2006; Sinner et al., 2004a) and mouse (Engert et al., 2013; Kanai-Azuma et al., 2002; Qu et al., 2008), however in zebrafish, *sox17* was duplicated resulting in *sox32* (Voldoire et al., 2017). The effect of the duplication on the development of endodermal structures in zebrafish is not well understood, and the reciprocal roles of Sox32/Sox17 TFs are still unknown, particularly the downstream set of endodermal genes that depends on the activity or repression of these two proteins still remains to be addressed.

3.2 The Sox family of TFs

Sox (SRY-related High Mobility Group Box) genes are an evolutionarily conserved family of TFs throughout the animal kingdom and are found both in vertebrates and invertebrates (Bowles et al., 2000; Wilson and Dearden, 2008). The first member of the *Sox* gene family identified was the *Sex-determining region Y* (*SRY*) gene. It is located on the Y chromosome and its loss of function results in sex reversal in males of both mice and humans (Hawkins, 1993; Koopman et al., 1991).

The *Sox* genes are grouped into 11 subfamilies according to their phylogenetic relationship; 9 groups are found in mammals - groups A, B1, B2, C, D, E, F, G, and H - while groups I and J are found in frogs and roundworm (Bowles et al., 2000; Heenan et al., 2016; Kamachi and Kondoh, 2013; Nagai, 2001; Okuda et al., 2006). A total of 20 *SOX/Sox* genes with diverse functions are found in humans and mice and 27 *sox* genes are present in zebrafish (Figure 3.1). All *Sox* genes share a highly conserved DNA binding High Mobility Group Box (HMGB) domain (sequence identity >50 %) and can contain additional domains such as trans-activation, trans-repression or dimerization domains (She and Yang, 2015). The HMGB, which consists of three α helices (Xu et al., 2002), binds to the minor groove of DNA with a preference for the sequence WWCAAW, where W indicates A or T. This binding causes a bend in the DNA strand and can enhance the recruitment of other TFs, which then bind to regions adjacent to the HMGB binding sites (Harley et al., 1994; Struckmann et al., 2011).

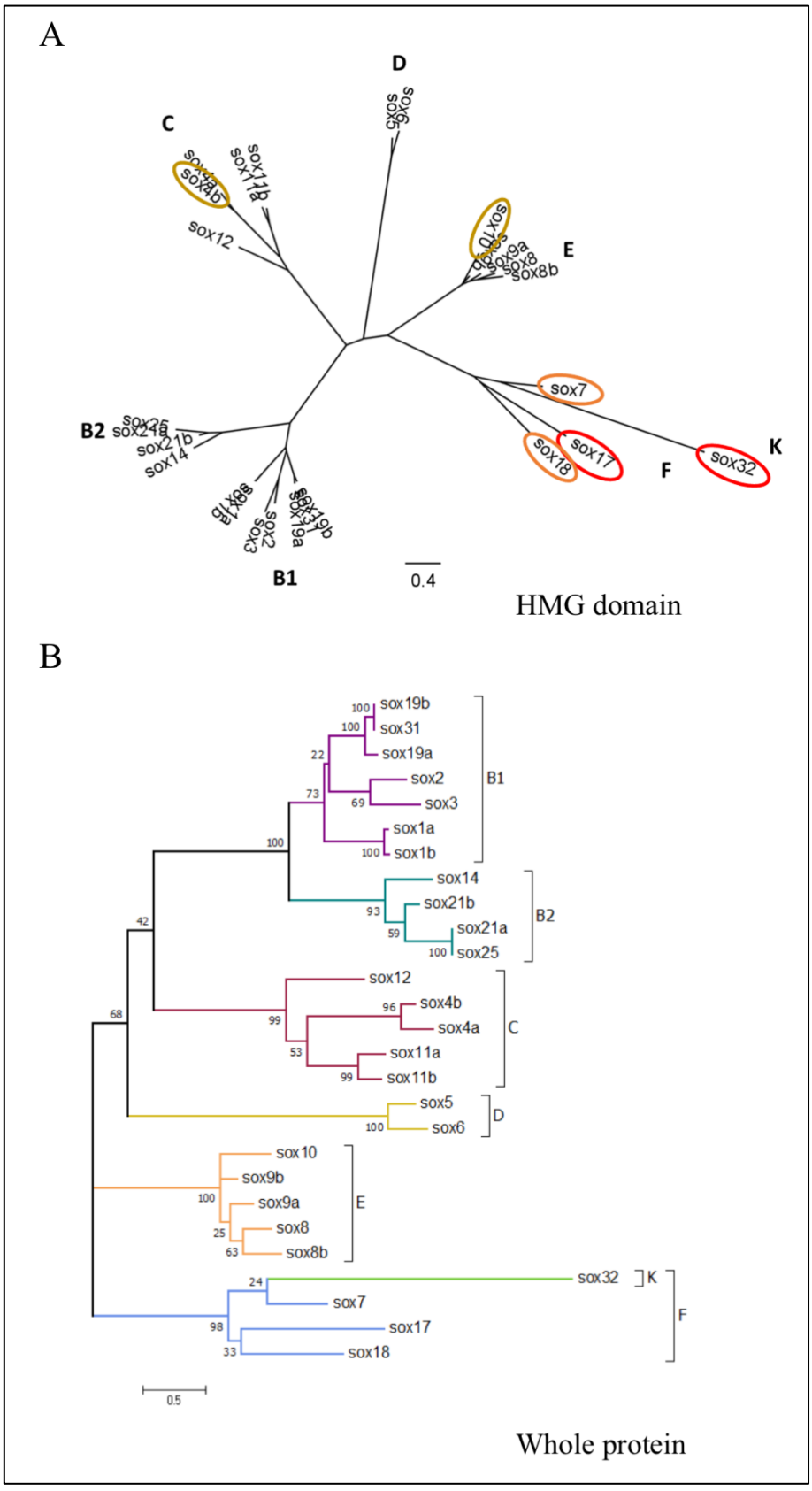


Figure 3.1 Phylogenetic tree of Sox proteins in zebrafish. Sox proteins have been grouped based on the sequence and structural similarity of their HMG box domain. (**A**) Phylogenetic tree rooted using only the highly conserved HMG domain in zebrafish. (**B**) Phylogenetic tree constructed using the whole sequence of all known members of *sox* family genes in zebrafish. Numbers on each node are the bootstrap values in thousand

runs. The marker length corresponds to a 10 % sequence difference. Note how *sox32* is most closely related to *sox7* and in the same node of *sox17* and *sox18*. Adapted from Bowles et al. (2000) with minor modification. Alignment and classification of the Sox domains by Amanda Evans.

Despite the highly conserved sequence of the HMGB domain, Sox proteins are capable of binding a variety of target sequences, by partnering with other TFs, such as those with a POU domain (Kamachi and Kondoh, 2013; Remenyi et al., 2003; Wissmuller et al., 2006). The HMGB/POU/DNA interaction is responsible for the transcriptional activation of multiple genes, including the direct roles of SOX2 and OCT3/4 in the regulation of the *Fgf4* in mouse embryonic stem cells (Rizzino, 2009). Other partnerships include SOX2 and PAX6 protein complex inducing δ-crystallin minimal enhancer DC5 during lens development (Kamachi et al., 2001), MEF2C and SOX18 complex in endothelial cells and MEF2C and SOX10 interaction in melanocyte development (Agarwal et al., 2011; Hosking et al., 2001). Furthermore, in zebrafish endoderm development, Pou5f3 has been shown to physically interact with Sox32 to regulate the expression of *sox17* (Reim et al., 2004). In addition to binding with other TFs, some Sox subfamilies, such as SoxD and SoxE, contain a dimerization domain, which allows them to bind to targets as both a monomer and a dimer (Bernard et al., 2003; Stolt et al., 2006). Overall, studies of the HMGB domain have reported its ability to conduct a varied range of gene expression control through transcriptional activation, epigenetic silencing and mRNA processing both alone and through partnering with other TFs (Prior and Walter, 1996; Wegner, 2010).

Sox genes have important functions throughout embryonic development (Kamachi and Kondoh, 2013; Prior and Walter, 1996; She and Yang, 2015; Wegner, 2010). For instance, SoxF proteins, such as SOX7 and SOX17 are coexpressed in the primitive endoderm in mouse (Artus et al., 2011; Kanai-Azuma et al., 2002; Lewis and Tam, 2006). In later endodermal stages, SOX2 contributes to the development of foregut (Schilders et al., 2014). Furthermore, SoxF genes regulate cardiovascular and neuronal development (Chung et al., 2011; Francois et al., 2010). In mice, frogs and zebrafish, Sox17, Sox7 and Sox18 have roles in both cardiogenesis and vasculogenesis (Cermenati et al., 2008; Herpers et al., 2008; Matsui et al., 2006; Zhang et al., 2005; Zhou et al., 2015). Three *sox17* genes have been detected in *Xenopus*, *sox17 α1*, α2 and β which together play a vital role in endoderm formation (Clements et al., 2003). They are activated during mid-blastula transition and precisely define the endoderm domain through gastrulation and neurulation (Hudson et al., 1997; Zorn and Wells, 2009).

Overexpression of *sox17* increases endodermal gene expression and changes the developmental cell fate program (Clements and Woodland, 2000). Gene knockdown experiments by morpholino oligos reveal that all three genes together are needed for the correct completion of gastrulation (Clements et al., 2003) and their ablation not only stops gastrulation but also results in a reduction of endodermal gene expression and as well as changing the fate of cells (Clements and Woodland, 2000; Hudson et al., 1997). Maternal VegT and Nodal-like signals establish the cascade of endoderm specification by activating zygotic genes *gata4*, *gata5*, *gata6*, *mixer* and *sox17* (Afouda et al., 2005; Engleka et al., 2001; Osada and Wright, 1999). Known direct transcriptional targets of Sox17 include *foxa1* and *foxa2* while β-catenin is an identified transcriptional cofactor of Sox17 in marking out the endodermal territory (Sinner et al., 2004b). In zebrafish, Sox17 and Sox32 are expressed in endodermal cells and in the dorsal forerunner cells (DFCs), a group of noninvoluting cells that are located at the leading edge of the shield during gastrulation, that form Kupffer's vesicle (KV), a ciliated fluid filled sphere with a vital role in laterality determination (Essner et al., 2005). In *sox32* mutants, fewer DFCs and a defective KV are visible, and consequently the mutants exhibit left-right (L-R) asymmetry defects (Aamar and Dawid, 2010; Alexander et al., 1999).

3.3 Chromatin regulation in early embryonic development – the advantages of ChIP-seq

The major function of a TF is to recognise and bind to specific sites in the genome, recruit cofactors, and regulate transcription (Spitz and Furlong, 2012). The first action of a TF is to find and bind to DNA and chromatin immunoprecipitation follow by sequencing (ChIP-seq) allows the binding sites of TFs to be identified across entire genomes (Hoffman and Jones, 2009; Johnson et al., 2007). By using ChIP-seq approaches, the DNA sequence motif that is recognised by the binding protein can be computed; the precise regulatory sites in the genome for any TF can be identified; the direct downstream targets of any TF can be determined and the clustering of transcription regulatory proteins at specific DNA sites can be assessed (Furey, 2012).

Perhaps the most important contribution of ChIP-seq approaches, however, is in providing a 'population' analysis of protein-DNA interactions on a genomic scale. This has shown how individual TFs employ different mechanisms for gene regulation depending on the degeneracy of the binding site recognition motif, the presence of other colocalised TFs and the distance from the transcription start site. In many cases, the mechanism of gene regulation by a given TF is specific to each particular binding site (Farley et al., 2015; Reiter et al., 2017; Spitz and

Furlong, 2012). Only through analysis of the entire range of binding sites in the genome can some higher functional principles be discerned.

ChIP can be used to understand the functional organization of the genome and to study complex mechanisms that involve changes in epigenetic signatures, TF and cofactor binding, chromatin remodelling and chromatin structure. The principle of ChIP is simple: enrich for a fraction of the chromatin using an antibody specific to a DNA-associated protein of interest. This technology was rapidly combined with massively parallel short read sequencing (ChIP-seq) and offers high specificity and sensitivity in profiling protein-DNA interactions (Aday et al., 2011; Hoffman and Jones, 2009; Schmidt et al., 2009). The first step of a ChIP assay is cell fixation where proteins are crosslinked to the DNA and the chromatin is isolated. Cross-linking is usually accomplished using formaldehyde. The DNA with the bound proteins is extracted from the cells and is fragmented by enzymatic digestion or mechanical shearing into fragments of average length ~200-500 bp. DNA fragments that are crosslinked to the protein of interest are enriched by immunoprecipitation (IP) with an antibody that specifically binds that protein. Subsequently, the crosslinking is reversed (the DNA is separated from the protein) followed by DNA purification of the IP-enriched DNA. In this manner, the protein bound to the DNA is enriched relative to the starting material due to purification with a specific antibody, and the enrichment of specific sequences in the immunoprecipitated DNA indicates that these sequences were associated with the protein of interest.

Analysis of these regions of the chromatin can be performed either by qPCR or by sequencing. If sequencing is used, adapters are added to the fragments which are then amplified to create a library that is then size selected (200-500 bp), before it is subjected to high throughput sequencing to generate millions of short reads. The reads are then mapped onto a reference genome to allow localization of protein binding. Traditional ChIP-seq provides limited resolution for TF binding sites due to high nonspecific background noise (Kidder et al., 2011). An evolution of this technology, ChIP-exo or high resolution ChIP, was developed in the laboratory of B. Franklin Pugh to increase the resolution of the ChIP-seq data (Rhee and Pugh, 2012). This technique reduces ChIP-seq peak width (the regions of the genome where multiple reads align suggestive of protein binding) using an λ-exonuclease enzyme to digest the immunoprecipitated DNA fragments to eliminate extraneous DNA and increase binding site resolution to within 20-95 bp. The result is better resolution of TF binding sites and identification of the precise location of TF-bound DNA (Carroll et al., 2014; Lim et

al., 2015; McHaourab et al., 2018; Pugh et al., 2018; Serandour et al., 2013; Starick et al., 2015; Wang et al., 2014). In addition, background signal is reduced due to the removal of excess DNA that is not bound by the protein of interest (Figure 3.2).

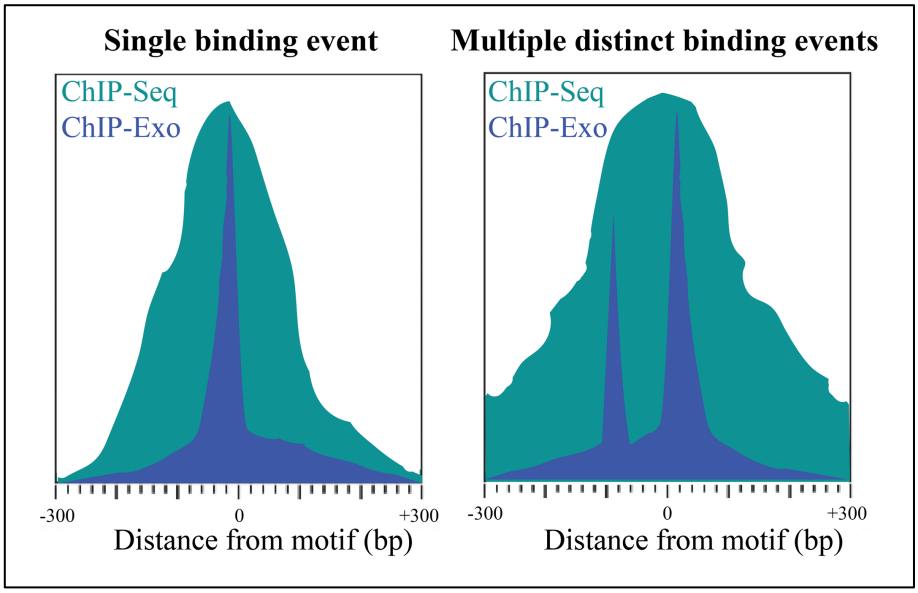


Figure 3.2 Advantages of ChIP-exo. Sharper peaks allow the identification of TFs binding sites at a resolution of 20-95 bp and distinguish multiple binding events in close proximity (left vs right panel). ChIP-exo yields lower background and higher confidence in defining DNA motifs and provides the ability to detect more precisely genome-wide protein binding profiles.

A ChIP-seq/ChIP-exo data set consists of millions of sequence reads that are generated from the ends of immunoprecipitated DNA fragments. The quality of called bases differs between the reads (the end of fragments has lower quality) and thus low quality data should be discarded before further analysis since read errors in the sequencing can drastically affect the finding of peaks. With current technology, read lengths are generally between 35 and 150 bp, single end or paired end. Once reads have been aligned to the reference genome, read position is used to infer binding site positions by applying peak detection algorithms to the mapped data. In this step, the aligned read data is transformed into a form that reflects the local densities of immunoprecipitated DNA fragments and the position(s) where the protein of interest was associated with DNA (known as peaks) is estimated. Visual inspection in a genome browser is also used to check positive/negative control sites for enrichment. Downstream analyses include *de novo* motif searching for the sequence of specific TFs, annotation of peak associated genes and gene ontology (GO) term enrichment analysis. Annotation is a very useful process in which the called peaks are linked to genomic information; possible target genes can be extracted by correlating peaks with promoter regions

or primary transcripts and the presence of specific peaks at the expected loci can be checked. The annotation can be further expanded with a GO/pathway analysis of the peak associated genes, uncovering biological processes and canonical pathways regulated through the targeted factor, thus discovering how a TF is involved in regulating a process either in the specific cell type or the whole organism. For many TFs, DNA-binding motifs have been published and can be searched for in the detected peaks, or *de novo* motif finding can be used to find overrepresented TF binding sites and the resulting motifs compared to published sequences. This powerful technique therefore helps us to understand the regulatory relationships between TFs and their target genes and allows us to establish models of regulatory regions that works across TF families.

3.4 Importance of antibody validation

In a ChIP assay, the success of the experiment is completely dependent upon the quality of the antibody used. For an antibody to work in ChIP, it must be used at the right concentration and it must be very specific, with no detection of non target proteins (Kidder et al., 2011; Landt et al., 2012). Antibodies can be raised in house or purchased from companies. In the case of the antibodies I used for my ChIP-exo experiments, I purchased them from a commercial antibody supplier, AnaSpec. In the first instance, at the beginning of my PhD, I attempted unsuccessfully to validate the antibody. I trusted the company that made the antibody and in the interest of time, proceeded to sequence the libraries before thoroughly testing and proving that the antibody was specific to my proteins of interest *in vivo*. I will first describe how I tried to validate the antibody at the beginning of my PhD, report on the final validation (radioactive immunoprecipitation assays) with associated evolutionary analysis done *a posteriori* and then describe the results from the bioinformatics analysis of my ChIP-exo datasets.

Antibodies anti-zebrafish Sox32, Sox17 and Mixl1 were bought from Anaspec, all of them were rabbit polyclonal. In particular, Sox32 antibody originates from a synthetic peptide derived from the C-terminal region of the protein, the one for Sox17 from the intermediate region of the protein. The company did not provide further information on the length of the sequence or the position as it was proprietary. For anti-Mixl1 the sequence was designed within the region 150-250 amino acids to avoid the highly conserved homeodomain region (amino acids 59-116). Two lots of this antibody were available JJ1505 and JJ1502. Anti-Mixl1 (JJ1505) was used before in a ChIP-seq experiment and shown to recognise Mixl1 (Nelson et

al., 2017), while both anti-Sox32 and Sox17 had only been validated by ELISA against the immunizing peptide by the manufacturer.

I began the validation process in the first instance by isolating protein from zebrafish embryos followed by western blotting (WB) using the same antibodies purchased for the ChIP-exo, anti-zebrafish Sox32 and anti-zebrafish Sox17. I was not able to observe a specific signal at the relevant molecular weight estimation of 35 kDa for Sox32 and 47 kDa for Sox17. Instead, bands were visible at around 50 kDa and 35 kDa with additional background signal (data not shown). This was likely because the secondary antibody was detecting the denatured primary heavy and/or light chains during western blotting. This caused a masking effect, making it difficult to detect my protein of interest.

Since I was not able to obtain satisfactory results, it could have been because that neither the Sox17 nor Sox32 antibodies were capable of recognising endogenously expressed proteins from zebrafish embryo lysates or low endogenous expression levels of Sox TFs in embryos at that stage (9.00 hpf) could have been responsible for the failed WBs, so I decided to test if the antibody could pick up larger quantities of *in vitro* translated protein. I generated protein for Sox32 and Sox17 using the *in vitro* translated protein system (rabbit reticulocyte lysate translation systems - RRL) and performed a WB against them with the anti-Sox32 antibody (Figure 3.3). This WB showed that the antibody was capable of detecting Sox32 at a low dilution of 1:200 of antibody.

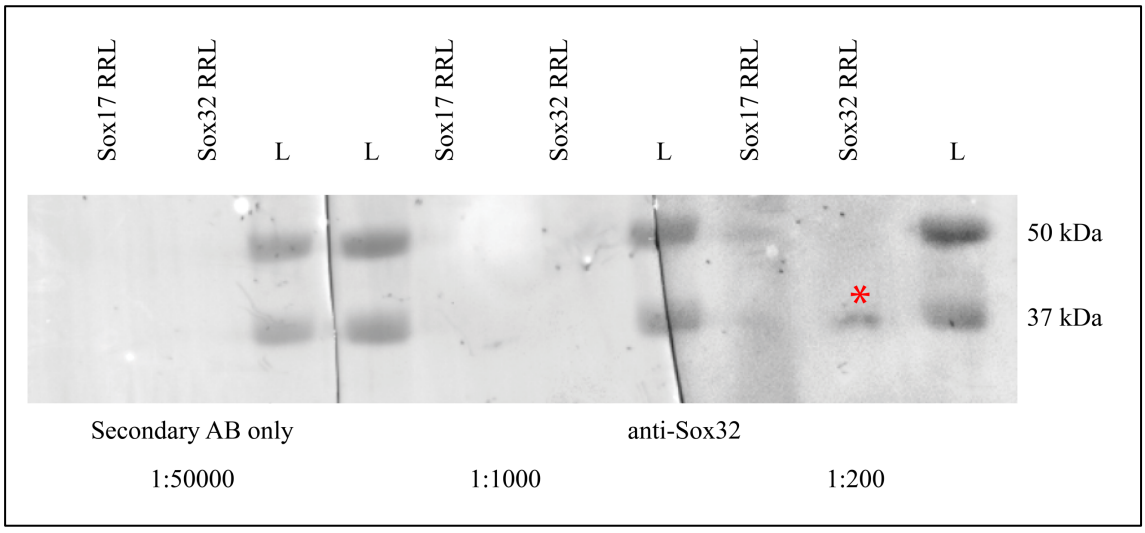


Figure 3.3 Western blot testing the binding affinity of anti-Sox32 antibody to *in vitro* translated protein for Sox17 and Sox32. Only incubation with the primary antibody diluted at 1:200 revealed a band for Sox32 (35

kDa) (red *), but no band for Sox17 (47 kDa) was observed, suggesting that the antibody was specific for Sox32. Staining without primary antibody results in no bands in the RLL. Molecular weight standards in kDa and antibody dilutions are indicated. L: ladder.

I then decided to test the antibody after overexpressing Sox32 in a non-zebrafish system, to see if the antibody recognised zebrafish Sox32 within the background mix of proteins of a mammalian cell. Sox32 is teleost specific, expressed in zebrafish and medaka (Alexander et al., 1999; Kobayashi et al., 2006) and has not been described in *Xenopus*, mouse and human, which is why I chose the human embryonic kidney cell line HEK293. I generated a plasmid where Sox32 was fused to GFP at the N-terminal (GFP-Sox32), so as not to interfere with the C-terminal epitope recognised by the Sox32 antibody. I then transfected three clonally different HEK293 cell lines with the fusion protein plasmid and confirmed expression of the protein with fluorescence microscopy (Figure 3.4).

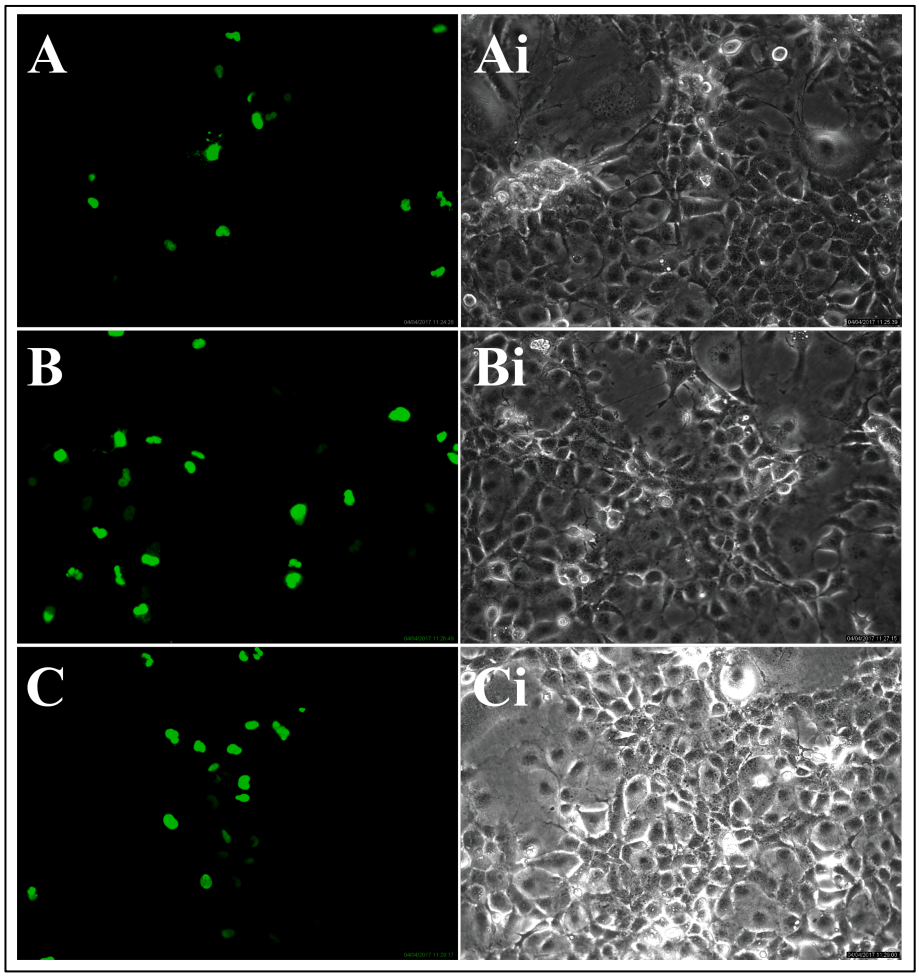


Figure 3.4 Transfected HEK293 cells with the GFP-Sox32 construct. (**A**), (**B**) and (**C**) Representative images taken with a 10x objective of three different HEK cell lines, confirming successful transfection with the GFP-Sox32 plasmid, successfully transfected cells are shown in green on the left panel. (**Ai**), (**Bi**) and (**Ci**) Brightfield

images on the right side.

I then performed a WB comparing lysates of HEK293 cells transfected with GFP-Sox32, untransfected HEK293 cells, HEK293 cells stably expressing eGFP and HEK293 cells stably expressing GFP, hence having two GFP controls (Figure 3.5). Immunostaining with the Sox32 antibody revealed a strong band for GFP-Sox32 (~56 kDa) in transfected cells (Figure 3.5, red *), however the blot also showed a band with a lower molecular weight in all four different groups – indicating that it recognised another protein, most likely a close family member with a similar amino acid sequence (Figure 3.5, black arrow). Staining the same samples with an anti-GFP antibody showed a band at the same size for the transfected cells (~56 kDa), suggesting that both the anti-Sox32 and the anti-GFP antibodies recognised the same protein (Figure 3.5, blue *). In contrast to the Sox32 antibody, the anti-GFP antibody did not recognise anything in the lysate from untransfected cells, and where the Sox32 antibody recognised an unspecific band in the positive control lysates, the anti-GFP antibody picked up only eGFP (~37 kDa) and GFP (27 kDa). It can be assumed that the most likely, unspecific target of the Sox32 antibody, lies within the mammalian family of Sox TFs. Considering the size of the unspecific protein (~45-50 kDa) as well as the high sequence homology of the targeted Sox32 epitope with mammalian SOX, it is highly likely that the unspecific protein being picked up was a member of SOX family. Transcriptomic analysis confirming the presence of SOX10 in HEK239 cells are currently unavailable, however SOX10 is expressed at high levels in human kidney – suggesting that its presence in HEK cells is feasible. Other family members (SOX2, SOX3, SOX6, SOX7, SOX8, SOX9) are expressed in HEK cells (The Human Protein Atlas).

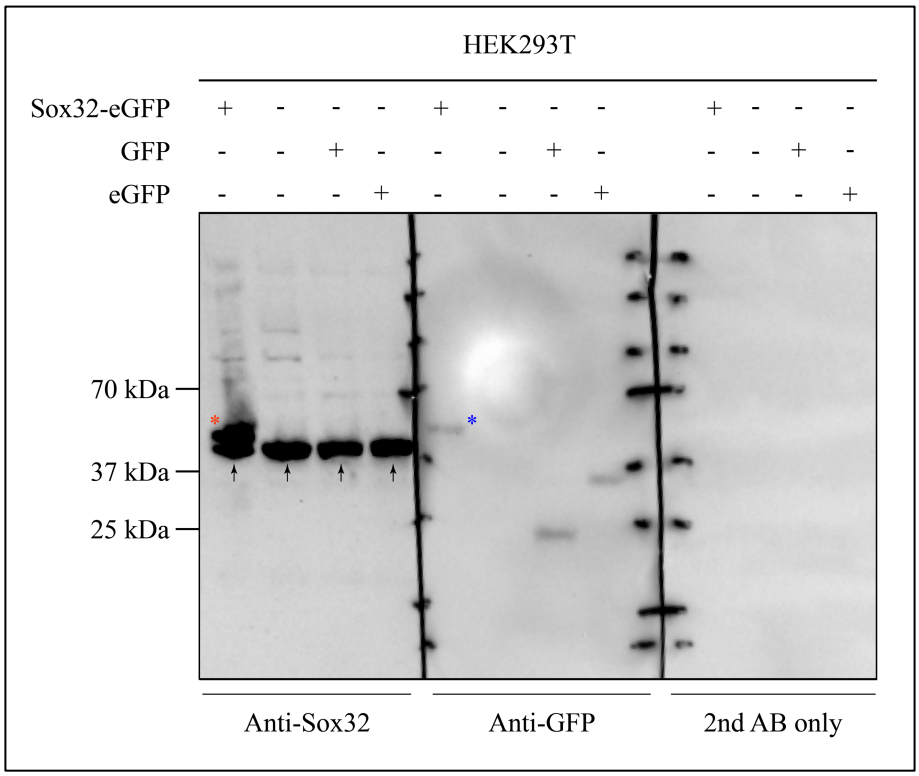


Figure 3.5 Unspecificity of anti-Sox32 in HEK cells. WB showing that the anti-Sox32 antibody recognised GFP-Sox32 in successfully transfected HEK cells (red *), but also picks up an unspecific, slightly lower band in transfected and untransfected controls (black arrow, left panel). Anti-GFP antibody was specific for GFP-sox32 (blue *), GFP and eGFP, and showed no band in the untransfected, GFP$^-$ HEK cells (middle panel). Staining without primary antibody resulted in no bands in any group, showing that the unspecific bands were not the result of the secondary antibody (right panel). Molecular weight standards in kDa are indicated.

To confirm this unspecificity hypothesis, the interaction between Sox32 antibody and other zebrafish Sox family members was investigated more closely. Radiolabelled proteins for multiple zebrafish Sox family members (Sox7, Sox10, Sox17, Sox18 and Sox32) as well as Mixl1 were generated in RRL by Amanda Evans. Mixl1, a protein with low sequence homology to Sox family members, was used as a negative control as it should not be recognised by anti-Sox32 antibody. IP with the generated proteins and Sox32 antibody conjugated beads was then performed (Figure 3.6). All the different radiolabelled proteins were observed in the input fraction (using X-ray film for detection), confirming successful protein generation. Sox7, Sox10, Sox32 and to a limited degree Sox18 were pulled down with the anti-Sox32 antibody, confirming that the binding of the Sox32 antibody extended at least to other members of the Sox family, and further confirming that the unspecific band seen in the previous WB might belong to other SOX proteins (Figure 3.5). Notably anti-Sox32 antibody did not recognise Sox17 protein.

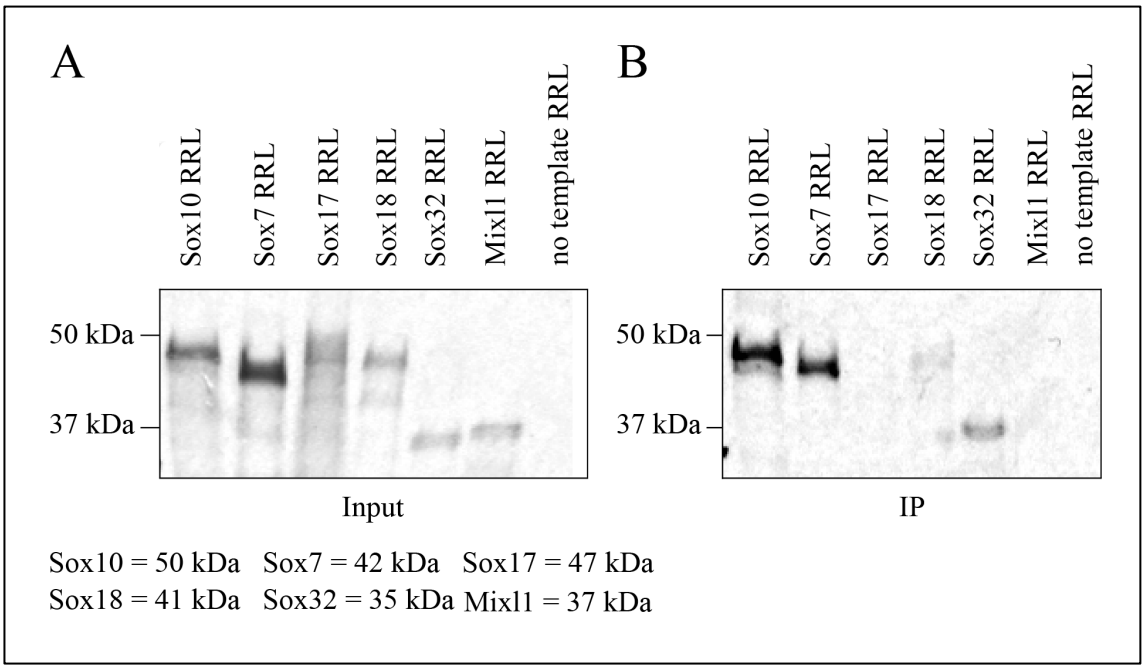


Figure 3.6 The rabbit polyclonal antibody used for anti-Sox32 ChIP-exo recognised Sox family proteins in radioactive immunoprecipitation assays. IP was performed using a radioactive methionine incorporated during protein synthesis in the RRL. **(A)** Detection of radiolabelled input fraction showed the successful translation of all Sox family members as well as Mixl1. (**B**) Immunoprecipitation of those proteins using anti-Sox32 rabbit polyclonal antibody followed by autoradiography revealed interaction of the antibody with Sox32 protein, but also unspecific interaction with Sox7, Sox10 and Sox18. Molecular weight standards in kDa are indicated. n= 2. IP performed by Amanda Evans.

The same IP experiment was performed using beads labelled with anti-Sox17 antibody. Similar to the Sox32 antibody, Sox17 antibody pulled down all the recombinant proteins except for Sox10, but including Mixl1, suggesting it was even less specific than the antibody raised against Sox32 (Figure 3.7). For this reason, no data from Anti-Sox17 antibody is described hereafter.

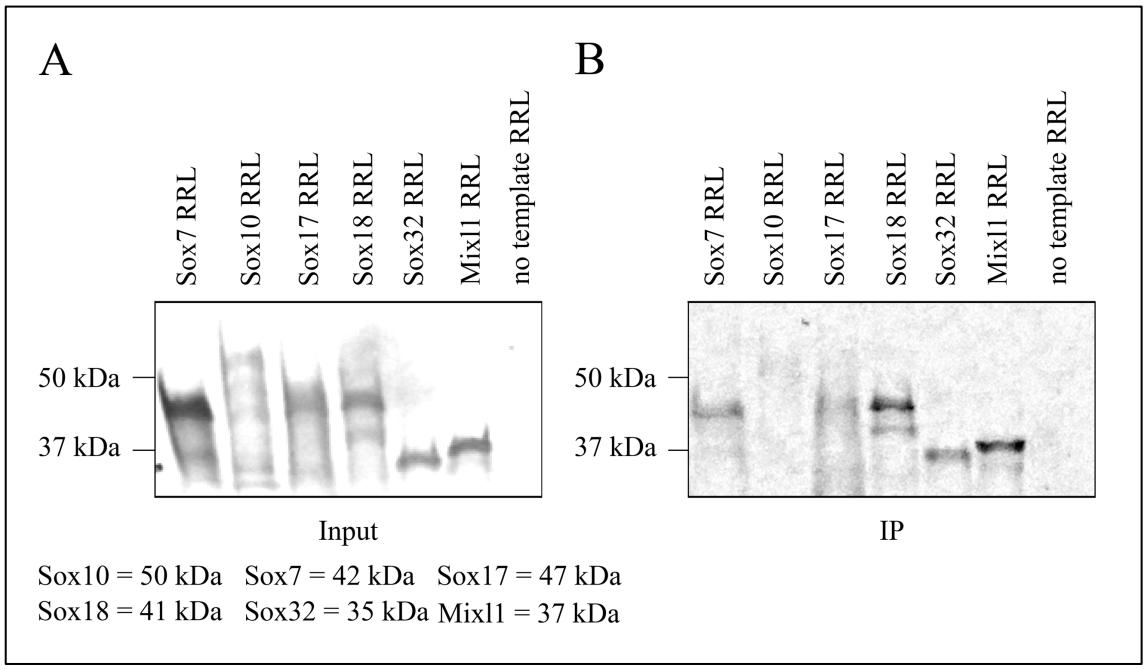


Figure 3.7 The rabbit polyclonal antibody used for anti-Sox17 ChIP-exo recognises Sox family members and Mixl1 proteins in immunoprecipitation assays. IP was performed using a radioactive methionine incorporated during protein synthesis in the rabbit reticulocyte lysate translation system. (**A**) Radioactivity incorporated was revealed by autoradiography and showed the successful translation of all Sox family members as well as Mixl1 in the input fraction. (**B**) Immunoprecipitation with anti-Sox17 labelled beads showed interaction of the antibody with Sox17 protein, but also unspecific interaction with Sox7, Sox18, Sox32 and Mixl1. Molecular weight standards in kDa are indicated. n= 2. IP performed by Amanda Evans.

Seeing that the antibodies raised against Sox32 and Sox17 were not specific for their respective targets, a further validation for the antibody raised against Mixl1 was conducted in the same manner (Figure 3.8). Both the two available different lots of the anti-Mixl1 antibody were able to pull down not only recombinant Mixl1 but also Sebox, the closest relative of Mixl1 (Figure 3.8,A); in addition, anti-Mixl1 antibody did not recognise Sox32 or Sox17 proteins (Data not shown). This antibody was previously used in ChIP-seq experiments (Nelson et al., 2017). However, the authors only tested the specificity of the anti-Mixl1 antibody against Eomes, an important TF expressed in mesendodermal cells which belongs to the T-box family. Eomes did not cross react with Mixl1 in RRL, but as those two TFs are not directly related, further experiments comparing anti-Mixl1 to closer related proteins such as Mxtx factors (mix-type homeobox gene, also known as Mtx) would be important to confirm specificity. Mix-family homeodomains share 50% amino acid identity with the homeodomains of both Mxtx1 and Mxtx2 (Hirata et al., 2000).

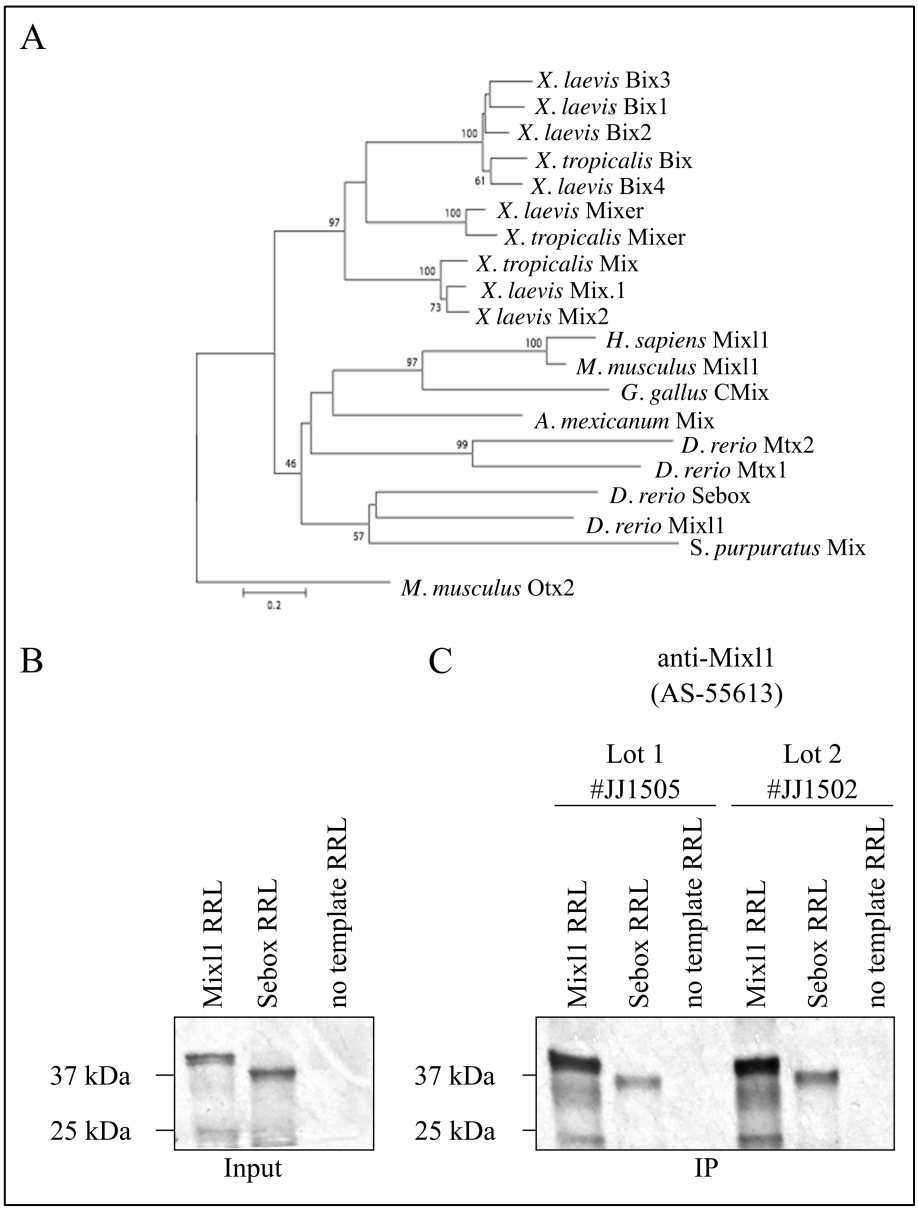


Figure 3.8 Anti-Mixl1 antibody recognises both Mix-like homeobox proteins Mixl1 and Sebox. (**A**) Evolutionary analysis of Mix genes in sea urchin, frogs, human, mouse and zebrafish. In zebrafish, the closest protein to Mixl1 is Sebox. (**B** and **C**) The rabbit polyclonal antibody used for anti-Mixl1 ChIP-exo recognised Mixl1 (37 kDa) and Sebox (36 kDa) proteins in immunoprecipitation assays. IP was performed using a radioactive methionine incorporated during protein synthesis in the RRL system. (B) Detection of radiolabeled input fraction shows the successful translation of Mixl1 and Sebox (C) Proteins pulled down with two different lots of the anti-Mixl1 labelled beads revealed interaction of the antibody with both Mixl1 and Sebox. Molecular weight standards in kDa are indicated. n= 2. IP performed by Amanda Evans.

The aforementioned validation experiments were done after I already had prepared and sequenced the ChIP-exo libraries and suggested that both anti-Sox32 and anti-Mixl1 were specific for multiple transcription factor family members but not specific for their intended

target. I consider this issue further in the discussion section. Ultimately, taken together this information was crucial for the interpretation of the bioinformatics analysis.

3.5 ChIP-exo-seq quality control

As mentioned earlier, ChIP with specific antibodies combined with sequencing provides a global snapshot of the genomic location where the protein under investigation interacts with chromatin. However, high background signal and relatively low mapping resolution (300 to 600 bp distance) are common drawbacks of ChIP-seq. ChIP-exo is an improved version of ChIP-seq that considerably increases signal to noise ratio. Incorporation of a lambda exonuclease (λ-exo) digestion step in the library preparation workflow allows footprinting of the left and right 5' DNA borders of the protein-DNA crosslink site.

Prior to the realisation that the antibodies were not target specific as described above, I performed ChIP-exo on 5.25 hpf and 9.00 hpf zebrafish embryos to detect Sox32 chromatin binding events *in vivo* with high resolution and on 5.25 hpf zebrafish embryos to detect Mixl1 binding events, using the aforementioned anti-Sox32 and anti-Mixl1 polyclonal antibodies. Preparation of ChIP-seq/ChIP-exo libraries wad a multistep process; several quality steps were needed throughout the workflow that involved both sample preparation and computational analysis, to provide robust experimental results and interpretation. I employed a number of quality control measures throughout the experimental protocol to ensure high quality sequencing results. The steps to obtaining a high-quality, non-biased DNA library can be divided into: i) having enough starting material ii) correct size of the sonicated chromatin and iii) suitable size of the libraries.

The high quality of ChIP-exo with less background signal relative to the standard ChIP-seq experiments comes at the expense of the amount of material needed for the input (He et al., 2015). The additional washing and digestion steps in ChIP-exo reduce the amount of DNA that is recovered prior to PCR amplification, but the recovery of substantial amounts of DNA is critical for the generation of a high-quality of the library. The number of embryos needed to reach a minimal amount of IP DNA is best determined empirically; as a starting point, a minimum of 500,000 cells containing the TF of interest is recommended (Gentsch and Smith, 2019). Previous ChIP-seq experiments at developmental stages similar to mine (Morley et al., 2009; Nelson et al., 2017; Nelson et al., 2014) used 5000 embryos per library; I used 9000 embryos per library in order to have enough starting material to reach a minimal amount of

immunoprecipitated DNA for ChIP-exo. Around 100-150 cells are marked by *sox32* WISH at 9.00 hpf, which totalled ~1,350,000 cells that should express Sox32 protein. This number was ~3x above the recommended number of cells and ~2x embryos per library.

Interactions between protein and DNA in the starting material is preserved by covalent crosslinking with formaldehyde. Two methods are routinely used to fragment chromatin, sonication (hydrodynamic shearing) and enzymatic digestion (micrococcal nuclease, MNase) and both methods preferentially fragment certain chromatin regions (Kidder et al., 2011). Sonication is preferred over MNase when using formaldehyde which limits the enzyme's activity (Haring et al., 2007).

Sonication conditions need to be optimised for each sample type and sonication instrument, with the goal of achieving fragmented chromatin in small fragments of 200-600 bp in length. Optimal fragmentation was determined by testing various sonication conditions on chromatin, followed by DNA isolation and monitoring of sonication efficiency by gel electrophoresis (Figure 3.9).

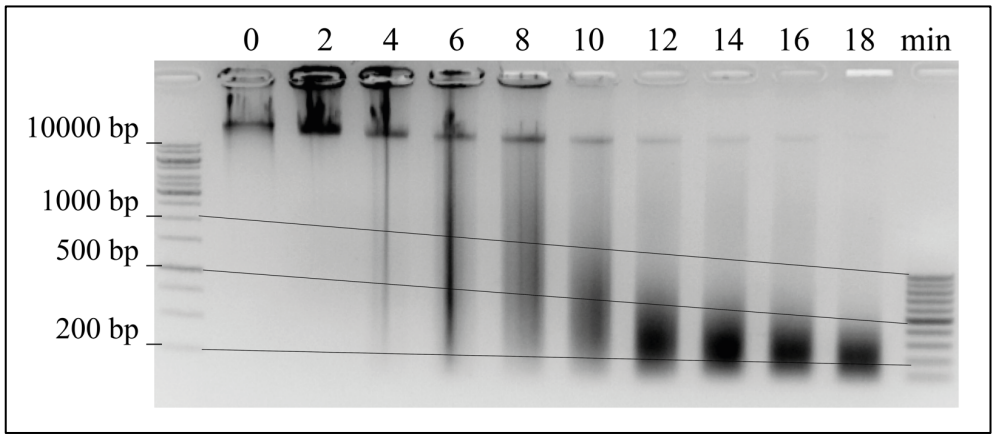


Figure 3.9 Time series to check chromatin shearing efficiency. Time series of sonication mediated shearing of chromatin-associated DNA as measured by gel electrophoresis. The image shows the results of gel electrophoresis of genomic DNA purified from crosslinked chromatin. Eight cycles of 30 sec On/30 sec Off (16 minutes total) gave the best range of DNA sizes (200-400 bp) with the majority of DNA fragments at ~300 bp.

Chromatin shearing is a time dependent process and my trials established 16 minutes (8 cycles of 30 sec On/30 sec Off) as optimal to obtain the desired fragment size. I therefore proceeded to fix 9000 embryos with formaldehyde for each ChIP-exo library and sonicated the chromatin under the aforementioned condition. The results are shown in Figure 3.10.

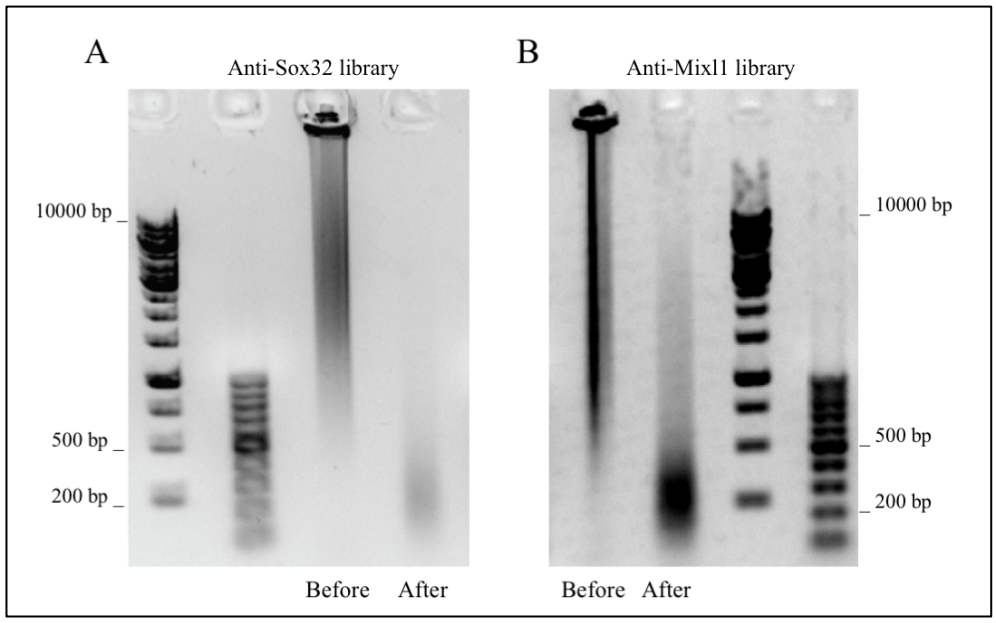


Figure 3.10 Successful chromatin shearing for ChIP-exo. (**A**) Anti- Sox32 library. (**B**) Anti-Mixl1 library. Samples were sonicated for 8 cycles of 30s On/30s Off. The crosslinks were reversed and the purified DNA was resolved on a 1% agarose gel. After sonication, DNA fragments fell inside the recommended range for chromatin immunoprecipitation (200-600 bp after sonication) and all high molecular weight gDNA (> 10.000 bp) used as input was fragmented. The DNA was then processed for antibody IP.

Before sonication, gDNA was visible above the 10 kb marker line demonstrating high-quality, non degraded template DNA. In optimal conditions, only one single, sharp band should be visible above the 10 kb marker for template DNA before sonication, in Figure 3.10 a smear is visible, that could be due to digestion of DNA by nucleases or/and shearing due to pipetting. Following sonication and decrosslinking, the gDNA was run on an agarose gel and all fragments fell in the range 200-500 bp, indicating that fragmentation had occurred. Efficient shearing of chromatin is important, because the resultant fragment size distribution determines the positional resolution of binding events.

After sonication and overnight IP with the appropriate antibody, the libraries were constructed according to the instructions of the Diagenode high resolution library kit (ChIP-exo protocol). At the end of the workflow, libraries were quantified both by fluorometry (Qubit) which uses DNA intercalating dyes and by qPCR (KAPA library quantification kits) which measured the number of molecules with successful incorporation of sequencing adaptors. Accurate library concentration is needed in order to avoid overloading the sequencing machine and overclustering.

Lastly, prior to sequencing, an important quality check was confirming the size distribution of the adapter-ligated library and determining the presence of any remaining adapter dimers (120 bp). If adapter dimers are present, they hybridize to the Illumina flowcell with a higher affinity than the library fragments, reducing the number of usable reads for subsequent bioinformatic analysis. Successful library preparation yielded a size range (200-500 bp) without any other contamination (Figure 3.11).

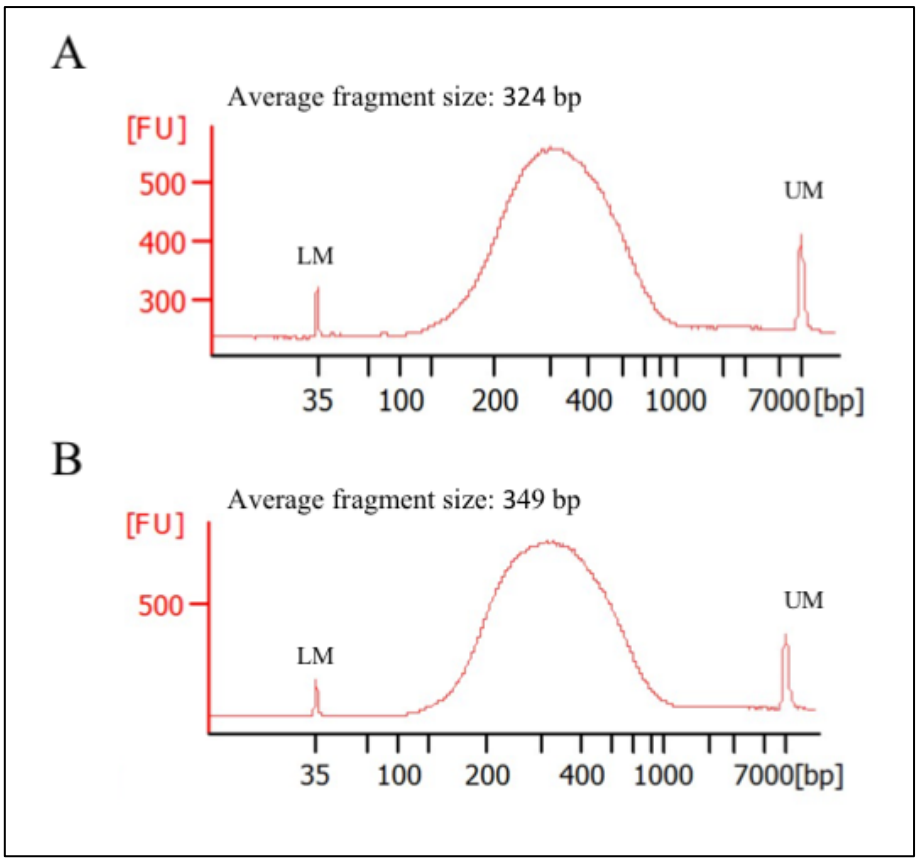


Figure 3.11 Library fragment size distribution. (**A**) Sox32 library and (**B**) Mixl1 library. Electropherogram of ChIP-exo libraries alongside ladder markers, that showed an enrichment of fragments at ~350 bp. LM: lower marker, UM: upper marker, FU:fluorescence units.

The quality and efficiency of library construction directly affects the integrity of the sequencing data; the aforementioned QC tests at multiple stages during library preparation helped to ensure the sequencing success of the libraries.

3.6 ChIP-exo-seq data processing and analysis

Libraries were sequenced using an Illumina HiSeq4000 sequencer as paired-end (PE) reads 50 nucleotides in length (P5 adaptors sequenced with R1 primers and P7 adaptors sequenced

with R2 primers); on average, 30 million reads were generated for each ChIP-exo data set. Despite great efforts to streamline the ChIP-seq procedure, no single ideal workflow exists, and there are many choices to select the relevant method for the data analysis. In addition, the major advantages of ChIP-exo are increased resolution of the peaks and the reduction of background noise, however, the currently available methods do not perform well in calling such sharp peaks. The real challenge is discerning between two close genomic regions as two distinct binding events and correlating each binding event to a specific gene(s). As a starting point, I checked the quality of the DNA sequencing itself, as this factor can directly influence the interpretation of the results. The base call quality for each raw sequencing data set was assessed using the FastQC program (Figure 3.12) (Andrews, 2010) and displayed as a box plot distribution at each base position. The results obtained for the quality score and base call distributions for R1 reads gave Q20 $\geq$ 95%, hence the majority of the dataset fell within the high confidence range (base quality score of 30-40, green region). In contrast, the average base quality score for reads from R2 was Q20 $\leq$ 80%. Not only did R2 reads show poor quality of sequencing but the program raised a warning for GC content proportion, sequence duplication levels and Kmer content representation. Reads coming from R2 were skewed towards high GC content and showed a high number of duplicate reads as shown in Figure 3.12 B and C respectively.

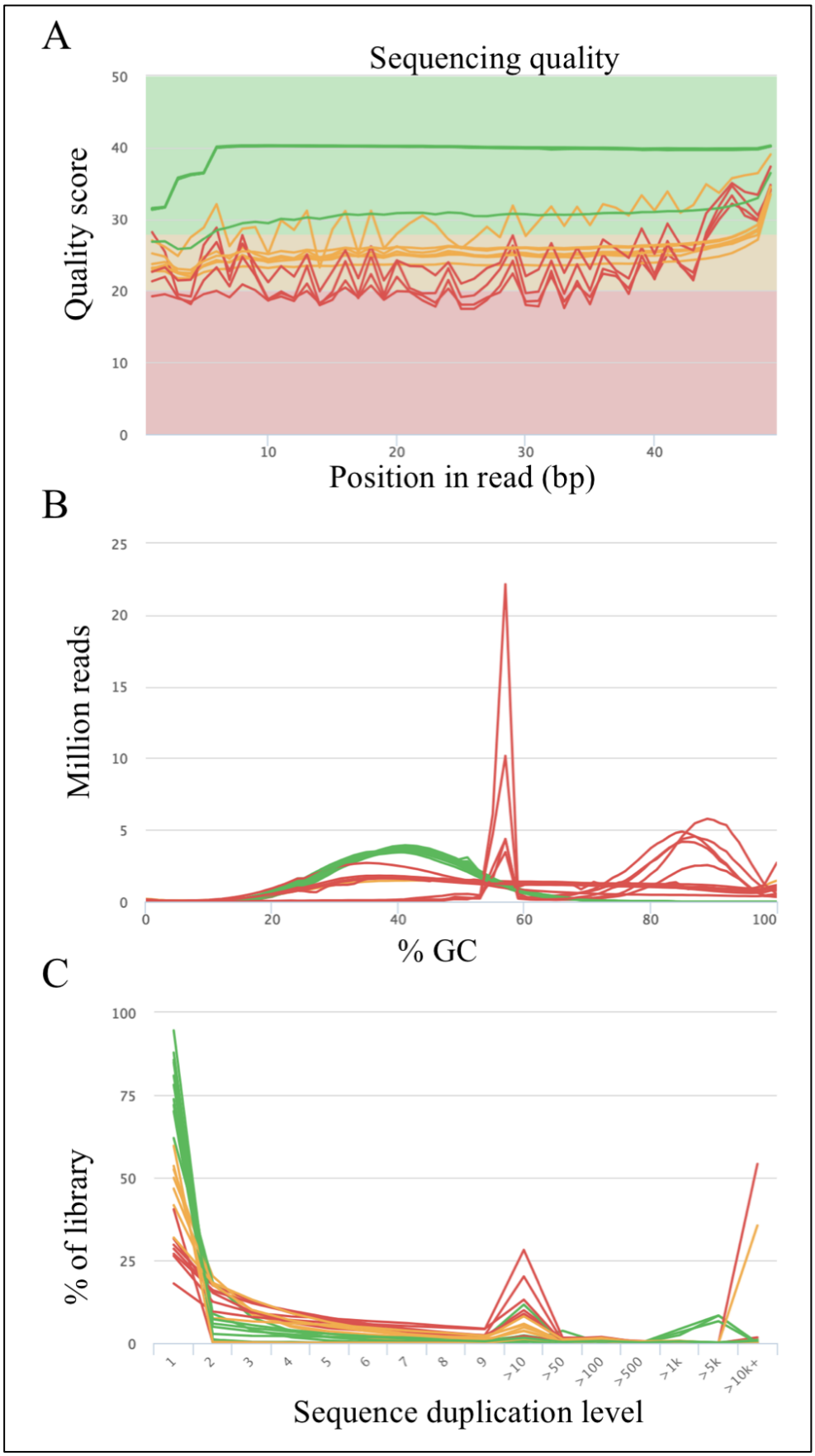


Figure 3.12 Sequence quality control for ChIP-exo data. (A) FastQC report of raw data illustrating the sequencing quality score. The y-axis on the graph shows the quality scores and the x-axis the position in the

read. The line indicated confidence in the base calls; the higher the score the better the base call and the green region designates a high confidence base call. All the green lines were R1 reads which pass the quality cut off. All the red and yellow lines were R2 reads, which did not sequence well. (**B**) Per sequence GC control plots showing the proportion of each base at each cycle. In a genome, all four bases are expected to be equally represented. Deviation from normal base content can indicate issues with library quality. All R2 reads failed this criterion as they contained high levels of GC (red line) whereas all R1 reads were in green showing that there was no difference in the proportion of bases in the sequenced reads. (**C**) Sequence duplication levels showed the high percentage of duplicated reads in R2 (red and yellow). Low degree of duplication were linked to R1 read (green). High levels of duplication are usually associated with enrichment bias (PCR over amplification); in the context of R2 reads they reflect the ligation step problem.

PE sequencing has been used previously for ChIP-exo libraries (Ye et al., 2016), however the authors used a modified version of the Illumina primers. Contrary to standard ChIP-seq library preparation, the ligation of P7 site adaptors uses an A-overhang base ligation; the same step in ChIP-exo is performed with a blunt-end ligation using T4 DNA ligase. Hence, the P7 site lacks a T base compared to regular libraries which becomes an issue for the pairing of Illumina primer R2 during PE sequencing. The Illumina R1 sequencing primer anneals to the P5 site adaptors, which contain the extra T and therefore single-end (SE) sequencing is not affected, explaining why all my R1 reads were of high quality. For the downstream analysis I decided therefore to use my libraries as SE and discard all the R2 reads.

Next, I aligned the R1 reads to the zebrafish Zv10 reference genome build using Bowtie1 (Langmead, 2010), retaining uniquely mapped reads and allowing 2 mismatch (-v 2 -m 1) and output alignments into SAM formatted files using -S option. The unmapped reads were trimmed by 6 bp from the 5' end and were mapped again and subsequently added to the downstream analysis. Because of the ambiguity of reads that align to multiple locations throughout the genome, I only retained uniquely aligned reads for subsequent analyses. The resulting sbam files were then sorted and indexed using Samtools (Li et al., 2009). The purpose of having indexed bam files is to be able to easily view the data in the genome browser.

I also considered ENCODE QC metrics for assessing the overall quality of the ChIP-exo sequences, described in detail at https://genome.ucsc.edu/ENCODE/qualityMetrics.html. The ENCODE consortium suggests the following metrics for the interpretation of ChIP-seq data and in assessing ChIP-seq enrichment quality: normalized strand cross-correlation (NSC) and relative strand cross-correlation (RSC). My ChIP-exo NSC and RSC values were close to the thresholds defined by the ENCODE consortium (NSC > 1.1 and RSC > 0.8) (Figure 3.13).

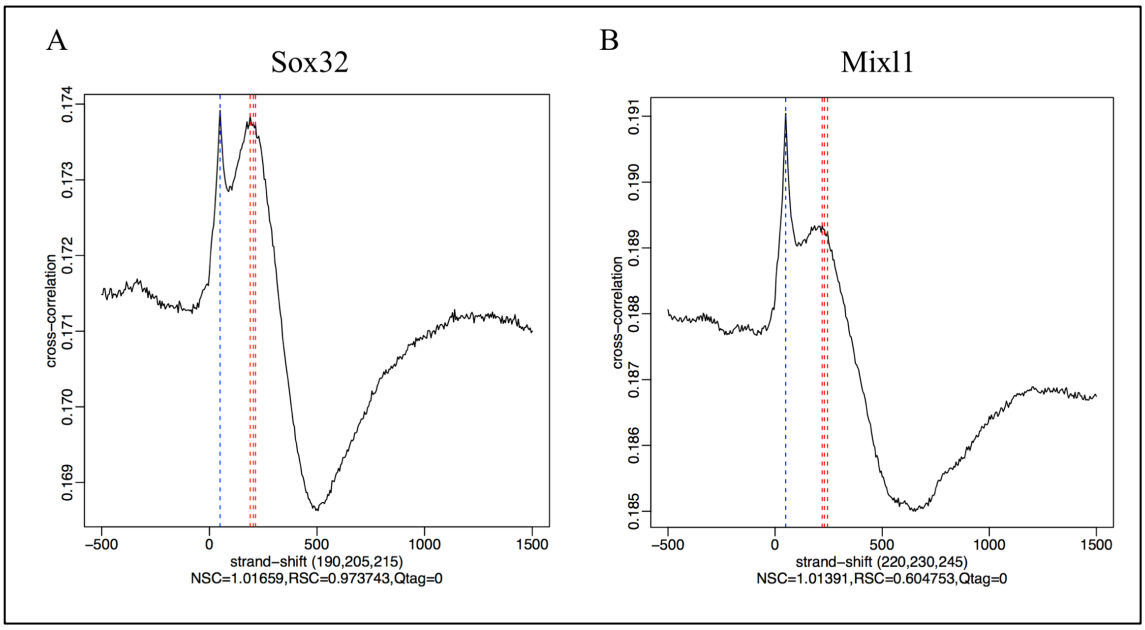


Figure 3.13 ENCODE quality metrics for ChIP-seq. Both Sox32 (**A**) and Mixl1 (**B**) libraries had values closed to thresholds endorsed by ENCODE. NSC and RSC values depict library complexity and how well the reads mapped to each strand are clustered around the locations of the protein–DNA interaction sites.

To avoid any incorrect interpretation of my sequencing data, I removed any redundant reads from the sequence alignment file with PICARD (https:// broadinstitute.github.io/picard). I then used irreproducibility discovery rate (IDR) analysis (Li et al., 2011) to identify the shared peaks in the duplicates at each time point. Lastly, similar to the pipeline described in (He et al., 2015) and (Tang et al., 2016), two algorithms were used with default significance cutoffs (GEM: FDR<0.01; MACS2 : FDR<0.01) to distinguish sets of highly significant peaks enriched on the genome. I then used 'findOverlaps' command from the Bioconductor package 'GenomicRanges' (Lawrence et al., 2013) to identify common binding events identified by both GEM (Guo et al., 2012) and MACS2 (Feng et al., 2011). BEDTools (Quinlan and Hall, 2010) was then used to create bigWig files to visualize common peak data. These analyses resulted in 9732 Mixl1 peaks and 9003 Sox32 peaks at 5.25 hpf and 5320 Sox32 peaks at 9.00 hpf after overlapping MACS2 and GEM (FDR<0.01). The observed reduction of Sox32 binding by 9.00 hpf is in line with the hypothesis that Sox proteins function becomes more specific as cell fate is restricted towards the end of gastrulation. These data also showed the highly dynamic occupancy of the *in vivo* genome binding patterns of this key TFs (Figure 3.14). At 5.25 hpf, Sox and Mix proteins shared the vast majority of peaks (5174) suggesting that these TFs frequently binds similar regulatory regions.

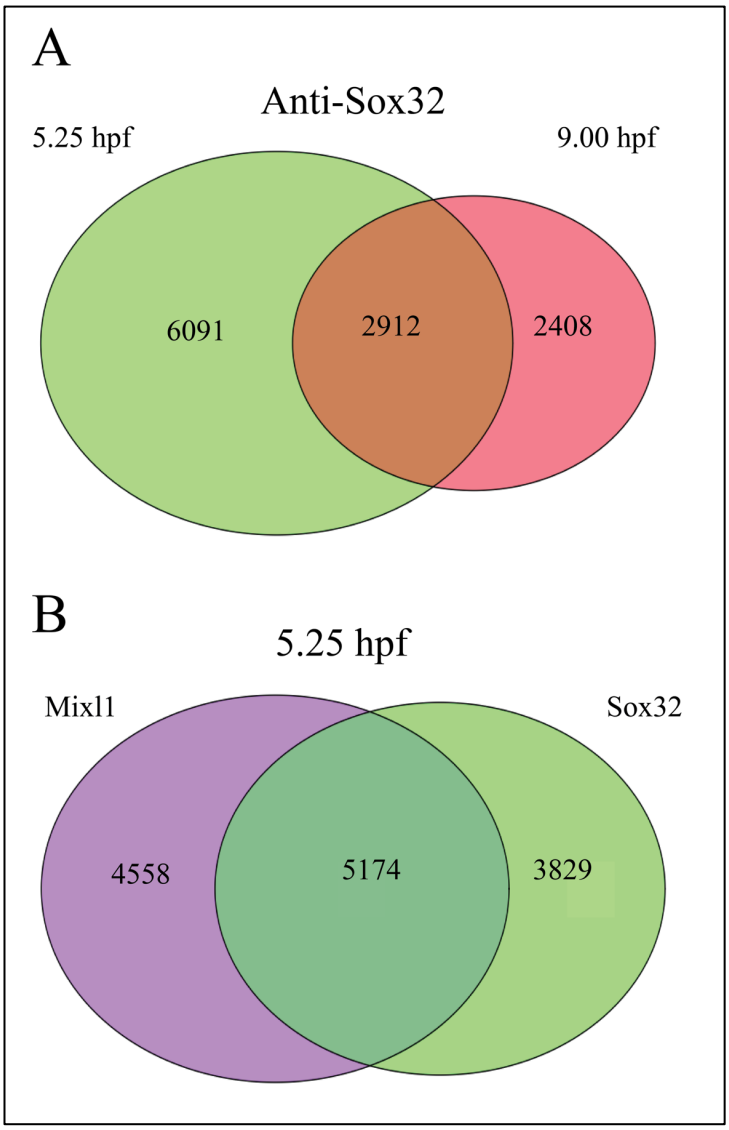


Figure 3.14 Mixl1 and Sox32 ChIP-exo peak distribution. (A) Venn diagram showing the overlap of Sox32 peaks between 5.25 and 9.00 hpf. Only 2,912 regions were shared between the 2 time points. The higher number of bound genomic sites (3,683) at early gastrulation stage may be associated with regulation of the activity of additional concomitant genes and could potentially indicate a broader role for Sox32 during mesendodermal induction. (**B**) Mixl1 and Sox32 share 5174 peaks at 5.25 hpf which may indicate that these common peaks are functionally relevant in the regulation of early mesendoderm genes.

Mapping and peak calling statistics are reported in Table 3.1 and for each merged biological library, a processed data file listing all the predicted binding sites common between GEM and MACS2 are provided in the Appendix.

Table 3.1 ChIP-exo statistics. Total number of reads, uniquely mapped reads and peak numbers are reported. M: millions of read.

Samples	# of total reads	# of uniquely mapped reads	% of uniquely mapped reads	# of peak summits identified (GEM)	# of peak summits identified (MACS2)	# of peak summits intersected (between GEM and MACS2), FDR <0.01
Mixl1 5.25 hpf rep 1	31 M	18,273,452	58.21%	22825	34865	9732
Mixl1 5.25 hpf rep 2	30 M	18,957,791	61.87%	21282	38127	
Sox32 5.25 hpf rep 1	35 M	26,502,388	75.13%	272343	33683	9003
Sox32 5.25 hpf rep 2	29 M	17,869,508	61.31%	31244	41025	
Sox32 9.00 hpf rep 1	27 M	17,834,944	65.61%	28139	40247	5320
Sox32 9.00 hpf rep 2	28 M	19,096,563	67.42%	32545	49705	

The advertised benefits of ChIP-exo (Rhee and Pugh, 2012; Serandour et al., 2013) over the previous ChIP-seq generation were higher resolution and the ability to map the small regions located around the actual protein binding sites using the power provided by exonuclease enzymes, ultimately providing greater confidence in calling peaks and identifying motifs. I therefore questioned how the ChIP-exo mean 5' tag density profiles looked in my datasets and whether I was able to sharply define boundaries of the TF location on the DNA. Reads aligned to the positive and negative strands shaped independent single peak summits for both Sox32 and Mixl1 ChIP-exo experiments. The sequences between the Mixl1 two binding site summits were sharper than the one in Sox32, which were 21 and 26 bp in length respectively (Figure 3.15). Thus, ChIP-exo-seq derived reads and the use of the λ-exo were able to sharply define the protein binding sites of the peaks.

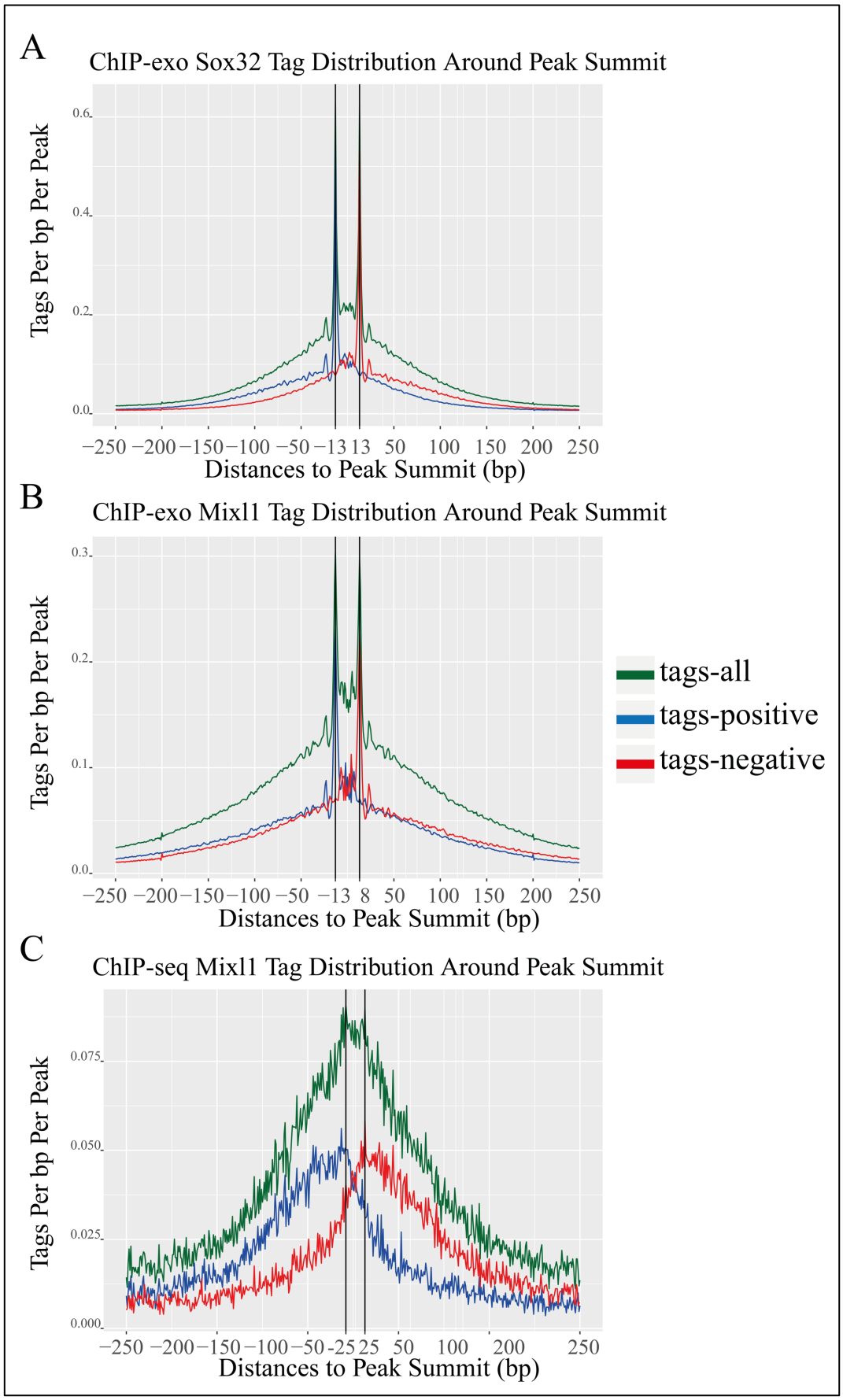


Figure 3.15 Analysis of Sox32 and Mixl1 *in vivo* footprints. Tag (5'end of read) distribution surrounding predicted binding events on average for Sox32, Mixl1 ChIP-exo libraries and Mixl1 ChIP-seq libraries (**A**)

Sox32 ChIP-exo and (**B**) Mixl1 ChIP-exo footprint showing 26 bp and 21 bp respectively as the distance between peaks on + and – strand. (**C**) Mixl1 ChIP-seq footprint from Nelson et al., (2017) showing how the + and - peaks are shifted away from the motif centre by 50 bp. Note the broad shouldering around the main ChIP-exo signal, probably due to incomplete λ-exo digestion. Tags from each DNA strand are plotted both separately (blue: +; red:−) and combined together (green).

Following the classic ChIP-seq pipeline, my next step was to annotate the genes closest to the peaks and determine whether I could associate these genes to specific GO terms, as gene set enrichment testing can enhance the biological interpretation of ChIP-seq data. I identified the genes associated with Sox32 and Mixl1 peaks by assigning peaks to their nearest genes within 10 kb of the gene body. However, the analysis was ineffective for both Sox32 and Mixl1 peaks and the overlapping peaks. In addition, *de novo* motif analysis was unable to identify any sequence closely resembling the previously described consensus binding motifs for Sox32 or Mixl1 within my ChIP-exo peaks. As shown in Figure 3.16 for Sox32 ChIP-exo at 5.25 hpf, 18.8% and 34.6% of the reads respectively were located in the gene bodies (introns and exons) and the downstream regions of a gene. Only 8% of the reads fell in the upstream regulatory region of genes.

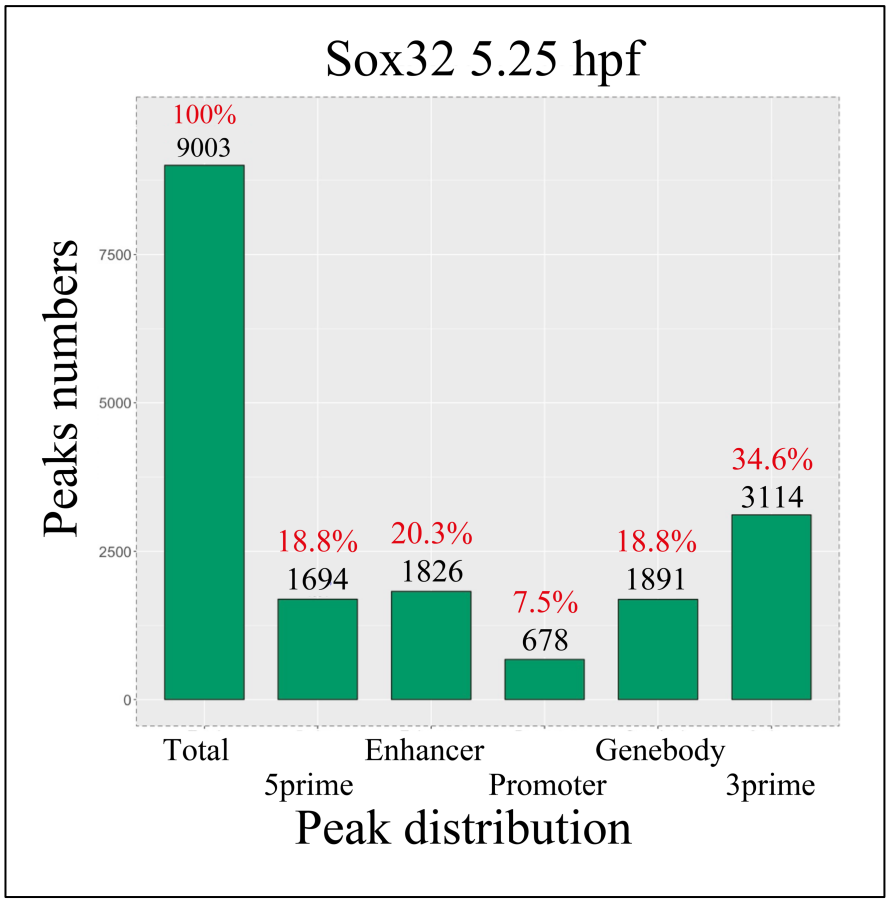


Figure 3.16 Genomic spatial distribution of all Sox32 peak summits at 5.25 hpf. The number of peaks

and relative percentage is shown. 5prime: 50 kb upstream of the TSSs. Enhancer: upstream region within −50 kb to −5 kb of TSSs. Promoter: −5 kb to +1 of TSSs. Gene body: from +1 of TSSs to end of transcripts. 3prime: 5 kb starting from end of transcripts.

An explanation for the observation of coverage in the gene body and 3' region can be found in the λ-exo treatment step of the ChIP-enriched DNA while the IP-DNA-Ab complexes remain on the magnetic beads. The λ-exo is predicted to stop its DNA digestion when it reaches the crosslinked protein, and this is the most critical step to achieve high mapping resolution. Conceivably, incomplete λ-exo digestion could lead to accumulation of 3' non-digested fragments, which in turn could result in fragments of different sizes and therefore low-resolution shouldering around peaks. This could also explain why no specific motifs were detected as the true motifs could have been obscured in the noise created by the non-digested fragments. Furthermore, the fact that both antibodies were family member specific and not target specific adds an additional confounding layer in the interpretation of binding peaks, and may explain why no motifs were enriched. TF family members can partly share binding sequences due to overlapping evolutionary origin but have different intrinsic binding affinity preferences which provide for their distinct *in vivo* functional specificities (Shen et al., 2018). Moreover, peaks could contain enriched motifs of non targeted TFs in addition to binding sites of the TF of interest, as has been previously described (Worsley Hunt and Wasserman, 2014). Finally, the ChIP-exo was performed on whole embryos and therefore it is possible that the obtained peaks are a mix of different TFs (i.e. Sox7, 10 and/or 18) expressed in cells not fated to become endoderm.

In contrast to traditional ChIP-seq protocols, ChIP-exo requires additional multiple sequential enzymatic reactions (end polishing, P7 exo-adapter ligation, nick repair, λ-exonuclease digestion, $RecJ_f$ exonuclease digestion) and the kit I used has been optimised and streamlined for mammalian systems. The ChIP-exo technique has been successfully applied to bacteria, yeast, mouse, rat, and human cells, however no record of *in vivo* work with *Xenopus* or zebrafish embryos is reported in the literature to my knowledge (He et al., 2015; Mahony and Pugh, 2015; Rhee and Pugh, 2012; Serandour et al., 2013; Starick et al., 2015). It is possible that the technique needs further optimisation in these model organisms.

I next sought to provide evidence of biological validity for the ChIP-exo data. I sought to validate the method using known endodermal genes and proceeded to inspect the upstream regions of Sox TFs (*sox32*, *sox17)*, Gata TFs (*gata4*, *gata6*) and FoxA TFs (*foxa2*, *foxa3*)

(Figure 3.17). I also analysed the relationship between my ChIP-exo signals and previously described ChIP-seq signals, in particular, I compared my Mixl1 ChIP-exo data to the published Mixl1 ChIP-seq data (Nelson et al. 2017). I identified multiple putative peaks in the proximity of these gene and proceeded to verify them by ChIP-qPCR (Figure 3.18). As shown in Figure 3.18, the integration of my ChIP-exo data with the published ChIP-seq data highlighted the functions of Sox32 and Mixl1 in regulating endodermal genes throughout gastrulation. Of particular note, where my ChIP-exo signals failed (*gata6*, *foxa3*) real peaks were detected by ChIP-seq; on the other hand, my ChIP-exo identified regulatory regions upstream of *sox17*, *gata4* and *foxa2* that were not detected by ChIP-seq.

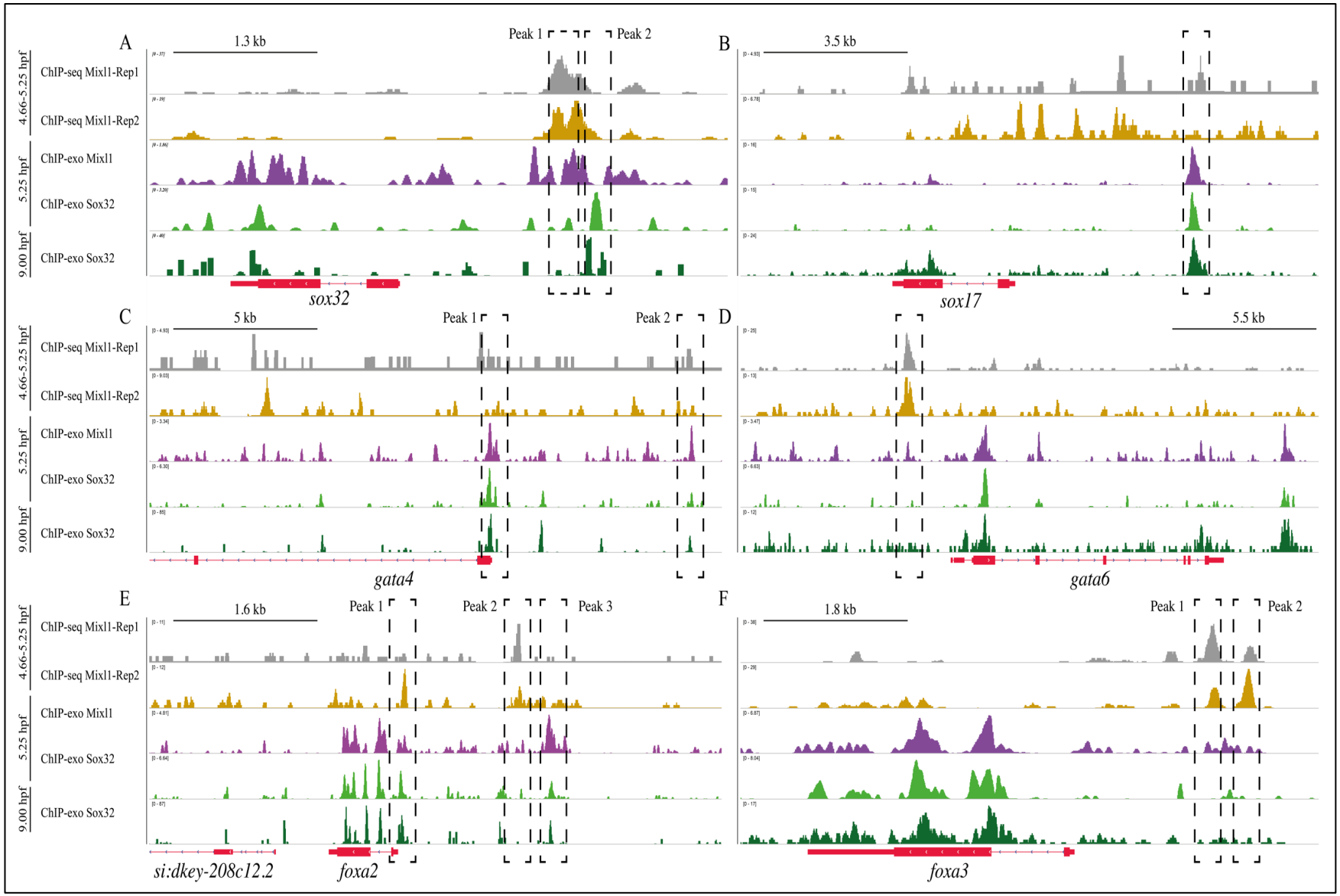

A
1.3 kb
Peak 1
Peak 2
ChIP-seq Mixl1-Rep1
ChIP-seq Mixl1-Rep2
ChIP-exo Mixl1
ChIP-exo Sox32
ChIP-exo Sox32
4.66-5.25 hpf
5.25 hpf
9.00 hpf
sox32
B
3.5 kb
sox17
C
5 kb
Peak 1
Peak 2
gata4
D
5.5 kb
gata6
E
1.6 kb
Peak 1
Peak 2
Peak 3
si:dkey-208c12.2
foxa2
F
1.8 kb
Peak 1
Peak 2
foxa3

Figure 3.17 Genome browser view of ChIP-exo and ChIP-seq signals for the indicated targets. Genome browser visualization of Mixl1 and Sox32 peaks at the target genes *sox32* (**A**), *sox17* (**B**), *gata4* (**C**), *gata6* (**D**), *foxa2* (**E**) and *foxa3* (**F**). Mixl1 ChIP-seq data from Nelson et al. (2017) are shown separated by replicate. ChIP-exo traces were generated from merged biological replicate pairs. Boxes represent peaks identified by common peak calls and which were then analysed using ChIP-qPCR. y-axis numerical values in each track indicate track height scaling in read depth.

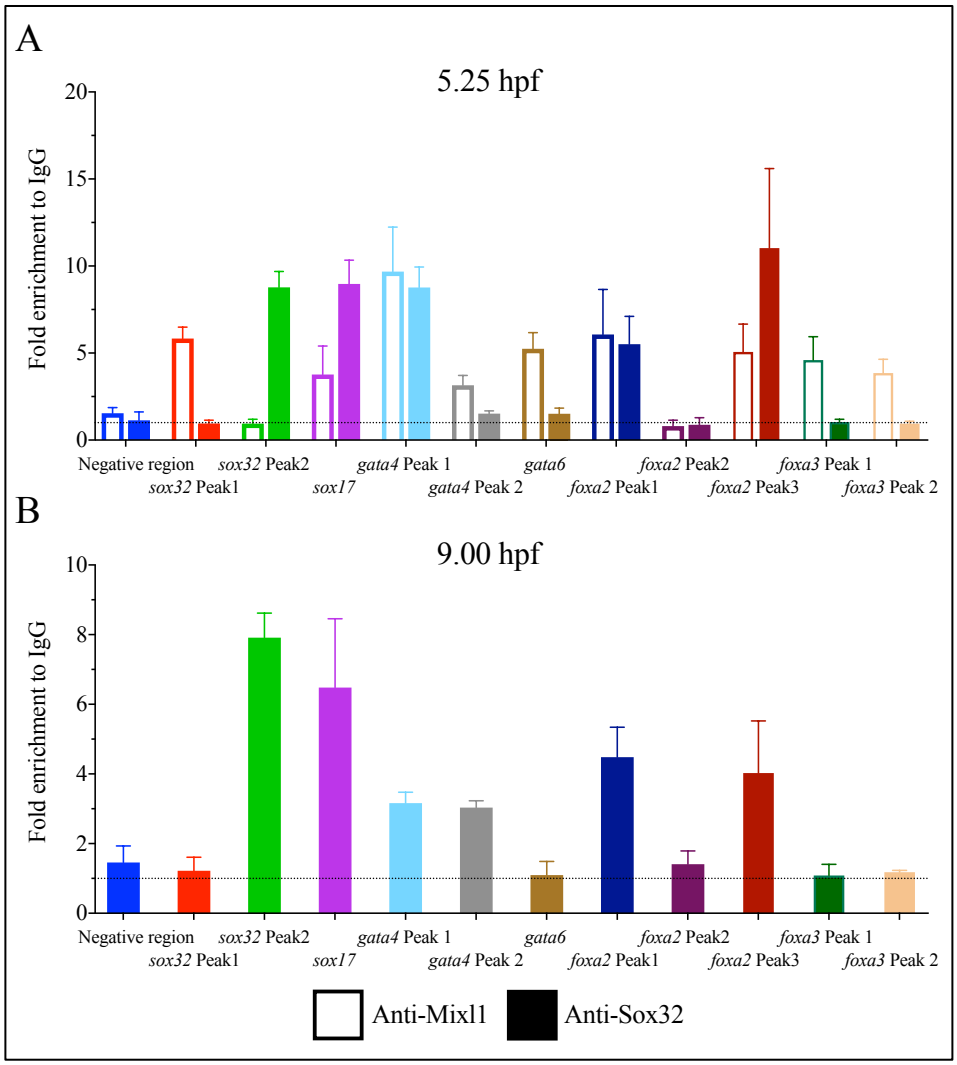


Figure 3.18 Mixl1/Sox32 ChIP-qPCR reveals direct regulation of endodermal genes during gastrulation. Anti-Mixl1 (empty boxes) or anti-Sox32 (full boxes) ChIP-qPCR of regions indicated in Figure 3.17 in embryos at 5.25 hpf (**A**) or 9.00 hpf (**B**). ChIP qPCR results showed that Sox32 and Mixl1 were efficiently enriched upstream of known endodermal genes. Two representative experiments are shown with mean fold enrichment over the IgG background ± SEM. Dashed line represents the related expression in the IgG control.

I then questioned whether I could use my dataset to add more direct relationships to the genetic connections proposed by (Tseng et al., 2011) in relation to *prdm1a* and *irx3a*. I also questioned whether I could confirm the role of Sox32 and Mixl1 in restricting the expression

of the mesodermal markers *vox*, *vent* and *gsc* in prospective dorsal endoderm and hence regulating mesendoderm specification along the dorsoventral axis (Imai et al., 2001; Perez-Camps et al., 2016; Sako et al., 2016). Lastly, as it has been previously shown that the pan-mesendodermal gene Tbxta represses the expression of *dusp6* (Morley et al., 2009) and synergistically with Tbx16 positively regulates *pcdh8* to establish directional cell migration of zebrafish mesodermal progenitors (Manning and Kimelman, 2015) I therefore questioned if these processes also depend on the activity of Sox32 and Mixl1. As shown in Figure 3.19, Sox32 and Mixl1 ChIP-exo/ChIP-seq identified multiple genomic binding events upstream of the above-mentioned genes during gastrulation, and the ChIP-qPCR signals (Figure 3.20) were relatively enriched in peaks upstream of the selected loci.


A
2.5 kb
ChIP-seq Mixl1-Rep1
ChIP-seq Mixl1-Rep2
ChIP-exo Mixl1
ChIP-exo Sox32
ChIP-exo Sox32
4.66-5.25 hpf
5.25 hpf
9.00 hpf
vox
B
Peak 1
Peak 2
4.7 kb
vent
C
1.6 kb
gsc
D
3.3 kb
sebox
E
2 kb
Peak 1
Peak 2
dusp6
F
5.3 kb
pcdh8
G
10.5 kb
prdm1a
H
3.5 kb
irx3a

Figure 3.19 Mixl1 and Sox32 bind mesodermal and endodermal genes. Genome browser visualization of selected Mixl1 and Sox32 peaks near target genes *vox* **(A)**, *vent* **(B),** *gsc* **(C),** *sebox* **(D),** *dusp6* **(E)** *pchd8* **(F)** *prdm1a* **(G)** and *irx3a* **(H).** Mixl1 ChIP-seq data from Nelson et al. (2017) are shown separated by replicate. ChIP-exo traces were generated from merged biological replicate pairs. The boxed enhancers represent peaks identified by common peak calls that were then analysed using ChIP-qPCR. y-Axis numerical values in each track indicate track height scaling in read depth.

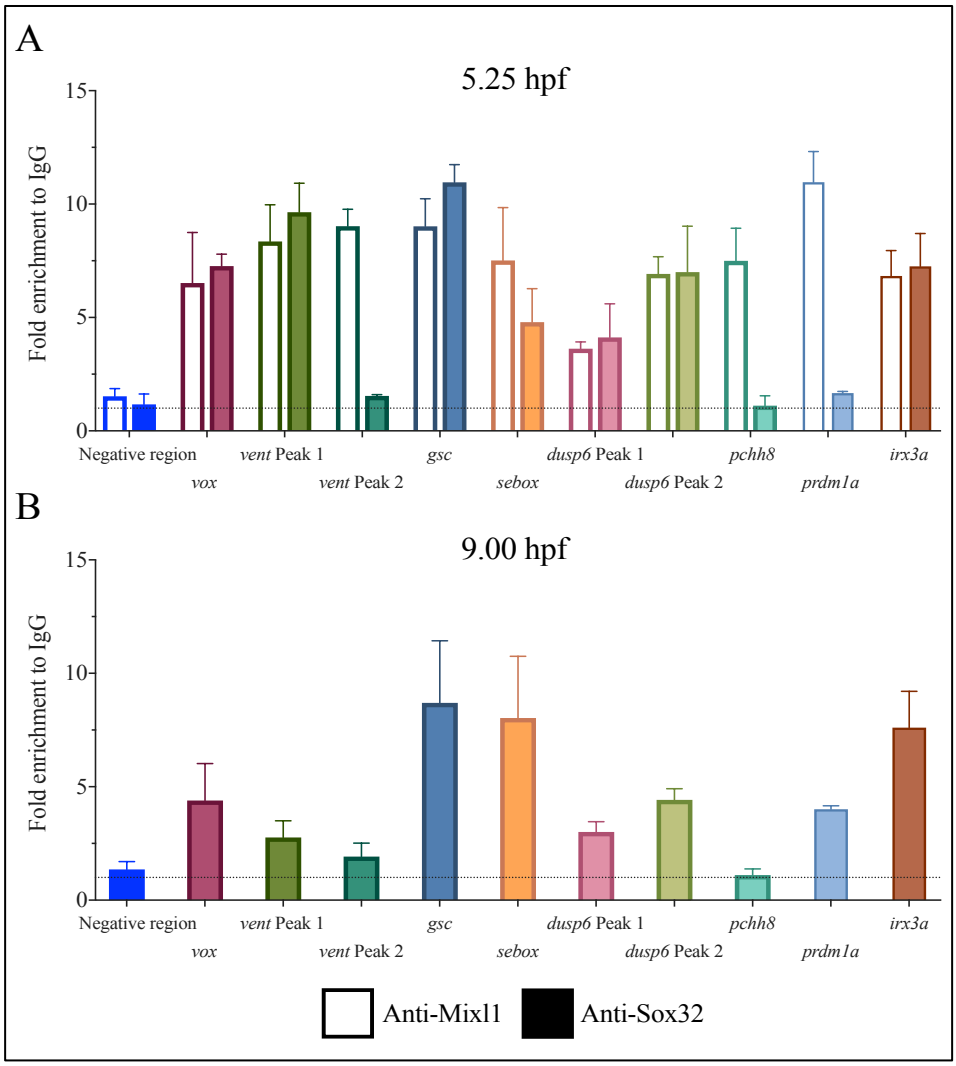


Figure 3.20 ChIP-qPCR revealed sites bound by Mixl1 and Sox32 proximal to novel endodermal and mesodermal regulators. Anti-Mixl1 (empty boxes) or anti-Sox32 (full boxes) ChIP-qPCR of regions indicated in Figure 3.19 in embryos at 5.25 hpf (**A**) or 9.00 hpf (**B**). Two representative experiments are shown with mean fold enrichment over the IgG background (dashed line) ± SEM.

Despite the limitations and caveats associated with ChIP assays and antibody specificities, by focusing only on previously known players of endoderm and mesoderm specification that have not previously being directly linked to Sox32 and Mixl1 regulatory function, establishment of novel interrelationships was possible. Collectively, these results highlight the

interplay of Sox32 and Mixl1 in establishing dorsal and ventral boundaries during gastrulation and activating expression of mesendodermal target genes and genes implicated in cell movements.

Taken together, these ChIP-qPCR results showed enrichment in the ChIP-exo datasets around endodermal and mesodermal genes and despite the antibody pitfalls, biological information can still be extrapolated from these data sets providing they are interrogated with caution, and any findings suitably validated for ensuring robustness of conclusions.

3.7 Discussion

ChIP is a powerful functional genomic technique to study mechanisms of gene regulation by enriching for DNA fragments that interact with a given protein of interest *in vivo*. Detection methods of interrogating ChIP-DNA fragments has shifted from detection of a single gene (ChIP-qPCR) to ChIP-on-ChIP (hybridization of the protein to DNA microarrays) to high throughput sequencing (ChIP-seq) to determine protein-bound (nucleosome-bound and TF-bound) regions of the genome.

The development of the ChIP-exo method has increased the genome wide scale, detecting DNA binding events with significantly higher spatial resolution, improved signal-to-noise ratio and increased sensitivity over the traditional ChIP-seq method (Mahony and Pugh, 2015; Rhee and Pugh, 2012; Serandour et al., 2013). The use of λ-exo to trim the 5′ end of each DNA fragment up to the point of each protein-DNA interaction boundary effectively footprints the transcription factor binding locations, resolving the positional organization of proteins within a complex, or uncovering alternative binding modes.

Regarding the fundamental disadvantages of ChIP-exo, not only does this technique suffer from the same limitations as ChIP-seq, including commercial antibody availability, specificity of the antibody and epitope accessibility, it also requires a larger number of cells and has a longer protocol comprising several enzymatic steps, making the methodology technically challenging to routinely adopt (He et al., 2015; Rossi et al., 2018).

Additional pitfalls in both ChIP approaches include optimisation of the crosslinking procedure, poor quality of chromatin sonication and the necessity of performing all steps on ice to prevent artefacts as far as possible. In addition, crosslinking is a reversible process and

as shown by Baranello et al. (2016), the crosslinking time can introduce strong bias in the protein-chromatin interactions. Thus, the duration of the crosslinking reaction and the timing and intensity of sonication must be empirically and carefully optimised (as shown in the results section) and the antibody needs to be carefully validated to avoid unspecific experimental outcomes.

The current guidelines on antibody validation for ChIP-seq are reported in (Landt et al., 2012), on the official ENCODE website (https://www.encodeproject.org/documents/ceb172ef-7474-4cd6-bfd2-5e8e6e38592e/@@download/attachment/ChIPseq_ENCODE3_v3.0.pdf) and are summarised in Wardle and Tan (2015) for the standards required for ChIP-seq experiments. The successful use of an antibody in experiments for specific applications such as western blotting or immunocytochemistry does not always correlate, validate or demonstrate selectivity in ChIP experiments. Hence, the ENCODE consortium recommends a two-step validation system: an initial immunoblot or immunofluorescence assay test, followed by at least one secondary validation test. Further validation can be carried out by performing western blotting or immunofluorescence on knock-out/knock-down systems or by running a pilot study to search for a known binding motif of the protein under peaks identified in the ChIP-seq data. Discovery of four-fold enrichment of the motif within the peak regions and motif presence in more than 10% of total peaks is also accepted as appropriate validation by ENCODE.

The quality of ChIP-seq results depends largely on the specificity of the antibody (how well it recognises and binds to the protein of interest). It is crucial to choose an antibody carefully, especially when antibodies are polyclonal and/or other members of the family of the protein of interest are coexpressed at the time point of interest, as the antibody could non specifically cross-react with these similar proteins. More and more companies are offering ChIP-seq validated antibodies, however those might not be available for the protein of interest or for the organism under investigation. In my case, no ChIP-seq validated antibodies for Sox32 existed, so I decided to investigate if the ELISA validated anti-Sox32 antibody (AnaSpec) could be used for a ChIP experiment.

The first step in validating an antibody was to ensure specificity in a WB. I initially tried zebrafish lysate (at 9.00 hpf when the gene is expressed), but I was not successful in detecting Sox32, possibly due to low overall protein expression levels and/or the low number of expressing cells. However, using the antibody against *in vitro* translated protein showed that

it could successfully recognise the protein. I tested the antibody against *in vitro* translated Sox17 protein to check it did not detect Sox17 in addition to Sox32; no signal was detected. However, these tests were quite limited in their informative value, as there is always the possibility of the antibody cross-reacting with any other given protein present in an organism. Unfortunately, this was not clear to me at the time and I proceeded to prepare the ChIP-exo libraries without additional antibody validation, planning to undertake a second validation in parallel to library preparation/data analysis in order to further confirm antibody specificity. The next logical step following western blotting and prior to library preparation would have been to ensure that there was no cross-reactivity with any closely related family member, especially as this particular antibody was raised against a conserved epitope, and the Sox proteins have highly conserved residues (Heenan et al., 2016; Kamachi and Kondoh, 2013; Wegner, 2010). At the beginning of my PhD I therefore made the mistake of trusting the antibody and proceeded to perform ChIP-exo with it.

Later, after realising the consequences of non-specificity of the Sox32 antibody, I wanted to understand if any relevant conclusions could still be derived from these data, and to do so I needed to understand what protein(s) the Sox32 antibody was reacting with. To investigate this issue, I first tested the antibody in HEK293 cells which do not natively express Sox32, by overexpressing Sox32-GFP. While the Sox32 antibody recognised the overexpressed protein, it also strongly recognised an unspecific mammalian protein at a slightly smaller molecular weight. Together with my colleague Amanda Evans, we decided to focus more on zebrafish Sox family members that share high sequence homology with the epitope against which the Sox32 antibody was raised. Previously published phylogenetic analyses show that Sox32 belongs to group F of the Sox transcription factors, together with Sox7, 17 and 18 (Chung et al., 2011). The closest related other group of Sox proteins is group E, containing Sox10, 9 and 8 (Bowles et al., 2000). Phylogenetic analyses performed in this study replicated these observations and we chose to test the specificity of the anti-Sox32 antibody against members of these groups, as we reasoned they were the most likely proteins to be recognised by the antibody. These tests were performed against *in vitro* translated protein of all other members of group F (Sox7, 17 and 18) and due to time constraints, against only one member of group E (Sox10). These experiments revealed that the anti-Sox32 antibody did indeed bind to Sox10, 7 and 18, highlighting how important performing phylogenetic analysis is in identifying unspecific binding partners for an antibody. Interestingly, the anti-Sox32 antibody did not recognise Sox17 protein. This analysis also led to the investigation of potential unspecific

interactions of the Sox17 antibody, which was shown to not only pull down family members (Sox 7,17,18,32) but also the completely unrelated Mixl1 protein.

If these phylogenetic analyses had been studied more thoroughly at an earlier timepoint in my PhD, I would not have performed the ChIP-exo with either the Sox17 or the Sox32 antibody. Rather, I would have relied on a tagged expression system (Lukoseviciute et al., 2018; Xu et al., 2012). Given the phylogenetic analyses and my own data, it is likely that the peaks identified in my Sox32 ChIP-exo experiments might not correlate to areas in the genome bound by Sox32, but to areas in the genome bound by Sox family members (7, 10 or 18). Sox 7 and 18 are important for vascular development not only in zebrafish but also in mice and *Xenopus* (Herpers et al., 2008; Zhang et al., 2005; Zhou et al., 2015). *sox7* is expressed throughout gastrulation and WISH experiments confirm s*ox18* transcript expression in lateral posterior mesoderm from bud stage (10.00 hpf) (Pendeville et al., 2008). In my work, I also showed that the Mixl1 antibody, which has previously been used in ChIP-seq (Nelson et al., 2017), is not entirely specific, and recognised family member Sebox which, similar to Mixl1, is an early target of Nodal signalling and regulates endoderm specification. Extensive testing for cross-reactions with the closest family members (Mxtx1 and Mxtx2) was not carried out and could be of interest for further investigations. Lastly, the results from the pull-down experiments with the Sox17 antibody prompted me to assume that any peaks derived from my ChIP-exo data using the Sox17 antibody could be a consequence of interactions with different proteins outside the Sox family, so I therefore decided not to attribute any clearly discernible biological value to the dataset.

Another drawback that I encountered analysing the ChIP-exo datasets was that, unlike ChIP-seq, there is no input control. Application of an input strategy helps in controlling the percentage of artefacts and increases confidence in peak calling in the standard ChIP-seq pipeline. In addition, the ChIP-exo protocol requires multiple ligation steps and an extra exonuclease step not present in ChIP-seq, which trims the left and right 5' DNA borders of the protein-DNA crosslink. This step inherently surpasses the limits of detection resolution of the average fragment size of a few hundred base pairs of ChIP-seq, however it also reduces the number of individual genomic positions. This increases the resolution of the peaks to which sequencing reads, however the available ChIP-seq algorithms do not perform well with such sharp peaks and the currently available ChIP-exo specific peak caller algorithms (Hartonen et al., 2016; Wang et al., 2014) rely on the existence of paired left and right exonuclease borders

to distinguish peaks. Furthermore, it is harder to distinguish peaks when large numbers of reads concentrate at a small number of bases, not knowing if they are due to over-amplification bias (PCR artefacts) or rather a real biological signal. Most importantly, I observed a high number of peaks located in the gene body and downstream region of genes that could be explained by different λ-exo digestion efficiency or ligation efficiency steps of the protocol; the presence of multiple enzymatic steps may mean that certain chromatin fragments are more prone to λ-exo digestion. Digestion bias might be partially explained by the fact that λ-exo activity is preferentially associated with AT rich sequences, the cleavage effectiveness being affected by the reaction temperature and as a further complication the bias is strand and nucleosome specific (Foulk et al., 2015; Meyer and Liu, 2014).

The datasets were therefore a mix of high-resolution TF bound regions and significant amounts of low-resolution shouldering regions, presumably from incomplete λ-exo digestion (Figure 3.17F, 3.19A,B,C). The exonuclease failing to stop at the crosslinking site or stopping at various points before the crosslinking site, explains not only the shouldering effect but also the high variability in peak width. The fact that Sox32 binds in several configurations and combinatorially interacts with different cofactors (Perez-Camps et al., 2016) could produce a fuzzy ChIP-exo signal. Additionally, the use of a protocol optimised for mammalian systems in zebrafish embryos may have resulted in sub-optimal library complexity, particularly in relation to exonuclease digestion time and subsequent fragment/peak size.

There are currently five ChIP-exo assay versions (ChIP-exo, ChIP-nexus, ChIP-tag-exo, ChIP-SSL-exo and ChIP-Exo 5.0); I use a Diagenode kit based on the original version, ChIP-exo. Each subsequent version has addressed and improved the technical limitations of the assay, in particular, reducing the input material and the level of adverse “shouldering” (undigested ChIP DNA) (Rossi et al., 2018).

Similar to Starick et al. (2015) and in contrast to Rhee and Pugh (2012) and Serandour et al. (2013), I found that only part of the ChIP-exo signal overlapped with ChIP-seq peaks and vice versa. The extra signal discovered solely by ChIP-exo may be due to the higher sensitivity of this assay in revealing individual binding sites, or instead may be false positives arising from non-specific peaks which would have been filtered out by the ChIP-seq input control, which the ChIP-exo assays lacks. Despite the described caveats (antibody specificity, exo-nuclease activity) I showed that some degree of biological significance could be derived from these data, which encourages me that my generated data may be of value.

In terms of the issue of non-specificity of the antibodies, as Mixl1 and Sebox act redundantly in endoderm formation (Kikuchi et al., 2000; Pereira et al., 2012; Poulain and Lepage, 2002), conclusions can still be drawn from the Mixl1 ChIP-exo providing one takes into account that the effects could be produced by either Mixl1 or Sebox. As these genes share a role in endoderm formation, conclusions pertaining to the endodermal GRN can still be drawn, providing these genes are considered together. Regarding the Sox32 antibody specificity issue, the other members of the Sox family that the Sox32 antibody pulled down were Sox7, Sox10 and Sox18. The only *sox* genes expressed in endoderm, and linked to endoderm formation, are *sox32* and *sox17*. Crucially, the Sox32 antibody did not pull down Sox17. *sox7*, *sox10* and *sox18* are expressed in different spatial domains and in different cells types and have never been linked to endoderm formation. I can therefore speculate that any binding events observed in my Sox32 ChIP-exo dataset that occur in the vicinity of known endodermal genes are as a result of Sox32 function, despite the lack of antibody specificity.

I leveraged ChIP-exo to confirm previously identified Mixl1 binding sites (Nelson et al., 2017) and to resolve novel Sox32 and Mixl1 genomic binding during endoderm specification, revealing new regulatory and functional relationships in the transcriptional hierarchy during early endoderm development in zebrafish and I have outlined new targets where Sox32 and Mixl1 act combinatorially to regulate gene expression.

Finally, I provide new insights into the previously described kernel of Gata factors described in (Tseng et al., 2011), outlining the important roles of both Sox32 and Mixl1 in redundantly activating endodermal transcription factors.

ChIP-exo mapping of Sox32 and Mixl1 should enable others to use these data sets for their own research to further understand the detailed interplay of Sox32 and Mixl1 in regulating the expression of cardinal endodermal genes in ultra-high resolution. In particular, in Chapter 6, I will describe how combining RNA-seq data with TF binding information gleaned from my ChIP-exo experiments can be used to inform the gene regulatory network governing zebrafish endoderm formation.

Chapter 4 – *sox17:GFP* transgenic line to study endoderm development in zebrafish

Chapter 4 highlights:

- In-depth characterization of the endoderm-specific reporter line *tg(sox17:GFP)* (Chung and Stainier, 2008), in which GFP is expressed under the control of the endodermal *sox17* promoter.
- Identification of GFP 'leaky' and 'non leaky' embryos, associated with both altered levels of gene expression and altered spatial expression patterns.
- Optimised protocols for live-cell dissociation followed by fluorescence-activated cell sorting (FACS) of early stage *tg(sox17:GFP)* embryos.

4.1 Introduction

Transgenic lines carrying fluorescent reporter genes have proven exceptionally valuable to the study of physiological processes and for cell lineage analysis, particularly in respect of transplantation experiments in animal models, including mice, frogs and zebrafish. Generation of green fluorescent protein (GFP) transgenic zebrafish exploiting tissue-specific promoters has provided valuable insights into gene regulation, organogenesis and morphogenesis, as the GFP fluorescence recapitulates the expression pattern of the targeted genes. Transgenic zebrafish where GFP protein was expressed under the control of an exogenous cis-acting element were generated first in the early 1990s (Stuart et al., 1988; Stuart et al., 1990). These early transgenes were generated with heterologous promoters that drove ubiquitous expression throughout the embryo, since then, many other lines have been generated that are driven by the promoters of specifically selected genes, leading to expression of GFP in subsets of tissues/cells and at specific timepoints. These transgenic lines fully exploit the external development and transparency of zebrafish embryos (Gong et al., 2001; Higashijima, 2008). Several transgenic lines driving fluorophore expression in endoderm derived tissue have been created using the promoters of *sox17*, *gata5, sebox, pou5f1* and *cxcr4b* (Chung and Stainier, 2008; Donà et al., 2013; Kikuchi et al., 2011; Parvin et al., 2008; Ruprecht et al., 2015). All these transgenic lines provide excellent experimental systems in which to start to dissect endodermal commitment during zebrafish development, however, as specified in the aims of the PhD, I was interested specifically in direct targets of the *sox32* gene, therefore in absence of a *sox32* reporter line I focused on the *sox17* transgenic line (a direct target of Sox32) to

further investigate the molecular mechanisms of endoderm specification. At the time of starting the PhD, *sox17* was the only validated direct target of *sox32*. Since its initial discovery in 1999 and the first attempts to molecularly characterise it, the *sox17* gene and its direct upstream regulator Sox32 have been recognized as critical components of endoderm development. However, precisely how the Sox transcription factors fine tune mesendoderm bifurcation and endodermal patterning has remained elusive (Aoki et al., 2002; Kikuchi et al., 2001; Ober et al., 2003). Two published transgenic lines were available that drive expression of a fluorophore under the control of the *sox17* promoter; both these stable lines were established in Didier Stainier's lab in 2008 (Chung and Stainier, 2008), by injecting one-cell embryos with plasmid DNA containing the same upstream sequence (approximately 4.2 Kb) of the zebrafish *sox17* gene fused to either GFP or DsRed. Transgenic founders were selected based on endoderm fluorescence/fish displaying specific expression of the fluorophore in the endodermal tissue. In this PhD I used the *tg*(*sox17:GFP)* line being already available in the lab. More recently, a new transgenic line harboring Kaede, a photoactivatable fluorescent protein expressed under the *sox17* promoter was developed *tg(sox17:Kaede*) by (Takada et al., 2018), which could prove useful in future studies of endoderm specification in zebrafish.

Green Fluorescent Protein (GFP), was isolated by Osamu Shimomura from the bioluminescent jellyfish *Aequorea victoria* in the 1960s for which he was awarded the Nobel prize in 2008 (Shimomura et al., 1962). The protein fluoresces green upon exposure to ultraviolet light, without the need for cofactors or enzymatic components. Fluorescent proteins such as GFP can be used as proxies for the study of protein expression and interactions. This is done by fusing the coding sequence for GFP directly with the coding sequence of the protein of interest. This can either be done *in vivo* by using DNA editing tools such as *tol2* transposon system or CRISPR/Cas9, or by generating a fusion protein *in vitro* and using it in overexpression studies**.** This allows the study of fluorescence (and therefore protein) distribution within the cell, monitoring of cellular migration and the functions of intracellular organelles and visualising protein interactions and gene expression in living cells. In particular, when the coding sequence of GFP is incorporated downstream of the regulatory region of interest, either in the genome or in a reporter vector it can be used as a reporter for the activity of promoter or enhancer sequences. Fluorescence intensity generated in such a system correlates with the activity of the regulatory region, and this can be used to investigate the interaction of transcription factors with suspected promoters (Gong et al., 2001). Thus, GFP has become an invaluable tool for studying biological processes in cells. A number of

optimised GFPs (eGFP) have been created from wild-type GFP by introducing point mutations that generate brighter fluorescence and a single, sharper excitation peak at 488 nm. In addition, a number of variants including blue FP, cyan FP and yellow FP have been produced by changing the spectral wavelength of GFP (Gong et al., 2001). The key benefits/strengths of using GFP as a genetic reporter are the protein is very stable (making GFP useful as a genetic tracer molecule being accurately distributed between newly divided cells); being a small molecule (27 kDa) the fusion of GFP to other proteins tends not to alter the function of the native protein and, compared to other fluorescent dyes, GFP does not generate free radicals when expressed in living cells, allowing for the study of dynamic and physiological processes. In addition, GFP continues to emit fluorescence even after fixation (Chudakov et al., 2010; Gong et al., 2001).

4.2 Characterisation of the *tg*(*sox17:GFP)* line

Firstly, to determine the ability of the *sox17*:GFP construct to drive faithful GFP expression in endodermal cells I compared *gfp* expression to *sox17* expression. Due to the lag time between transcription and GFP folding, I conducted RNA whole mount *in situ* hybridisation (WISH) against *gfp* in order to detect enhancer activity at different developmental stages. The expression was first observed as early as 5.25 hpf and by 9.00 hpf the salt-and-pepper pattern was recognisable (Figure 4.1).

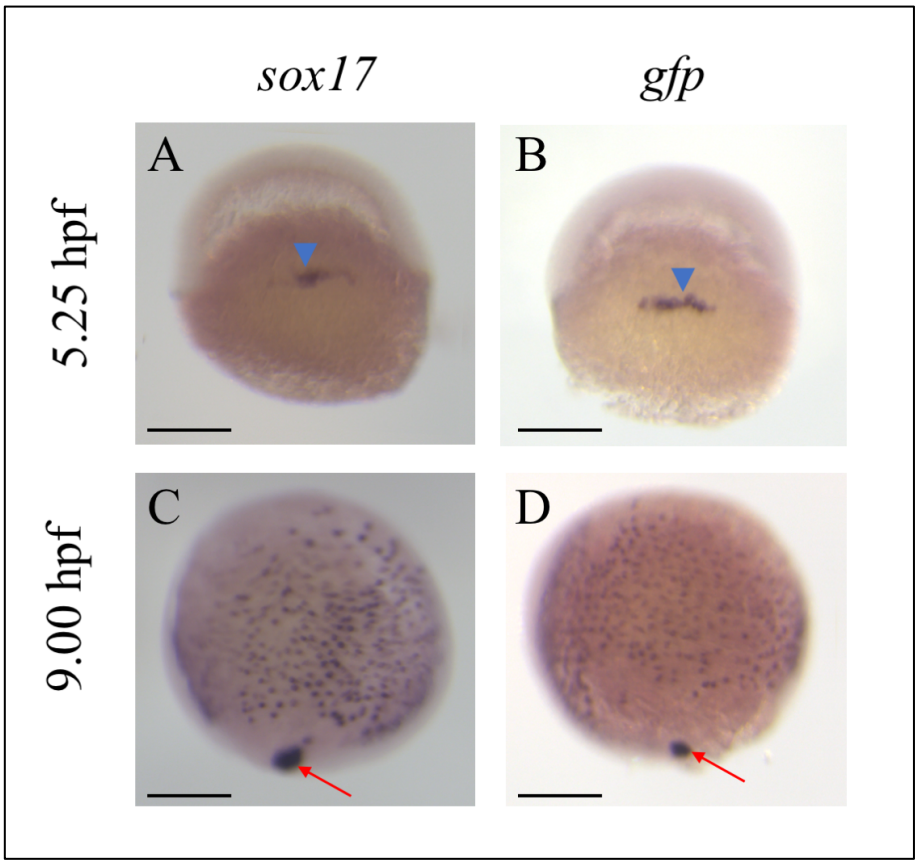


Figure 4.1 Expression pattern of *sox17:GFP* line. (**A**) and (**B**) Lateral views of *sox17* and *gfp* expression pattern respectively at 5.25 hpf. Both *sox17/gfp* signal marked the dorsal margin of the embryo (blue arrowheads). (**C**) and (**D**) as for (A) and (B) except at 9.00 hpf. Note the salt and pepper pattern of migrating endodermal cells at 9.00 hpf. Red arrow point to the Kupffer vesicle, a ciliated organ important for establishing left-right asymmetry in the embryos. Scale bars represent 250 μm.

The GFP expression patterns observed were similar to the endogenous *sox17* expression reported previously by WISH (Aoki et al., 2002 and Figure 4.38), suggesting that the 4800 bp sequence could drive *sox17*-specific GFP expression faithfully. However, unexpectedly, it was noticed that a minority of embryos were not able to recapitulate the expression pattern previously reported (from now, embryos showing the previously reported GFP expression are referred to as non leaky embryos). In particular, we observed embryos where GFP expression was not limited to endodermal cells (from now referred to as leaky embryos); rather, it was detectable throughout the whole embryo, suggesting incomplete or imprecise spatial control of the cis-regulatory elements of the *sox17* upstream region in this particular transgenic line (Figure 4.2 and 4.3). No temporal activation discrepancies were noticed amongst leaky and non leaky embryos, with both displaying detectable GFP transcripts at 5.25 hpf and no apparent phenotype was visible in leaky embryos at 24 hpf. More specific or more stringent assays may be needed to reveal a phenotype (Figure 4.3).

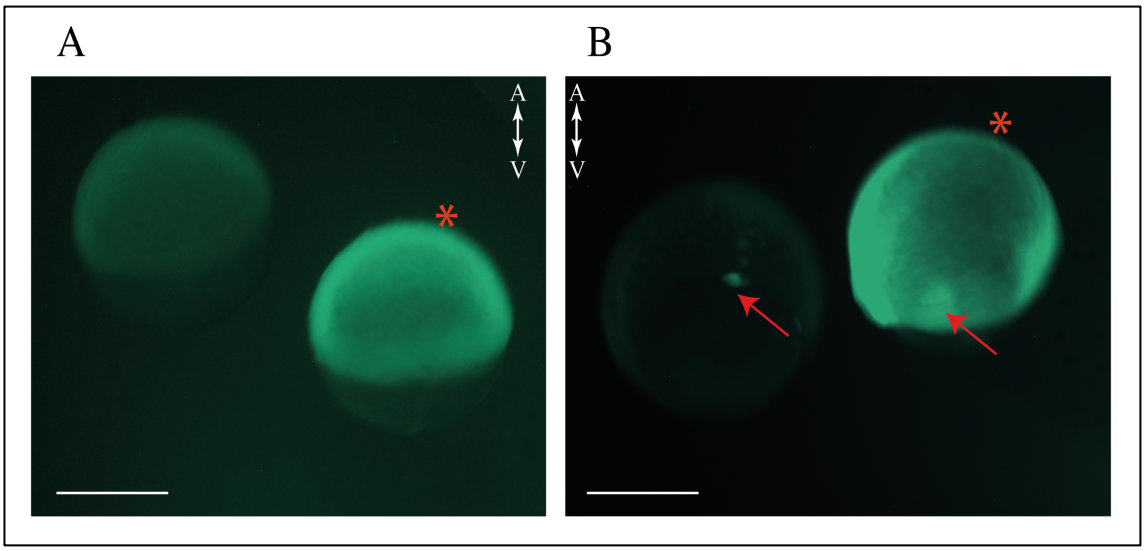


Figure 4.2 Non leaky and leaky *sox17:GFP* embryos. (**A**) Lateral views of native GFP expression in *sox17:GFP* embryos at 5.25 hpf (**A**) and 9.00 hpf (**B**). Red * denote leaky embryos, red arrows point to the Kupffer vesicle. A: animal pole. V: vegetal pole. Scale bar represents 250 µm.

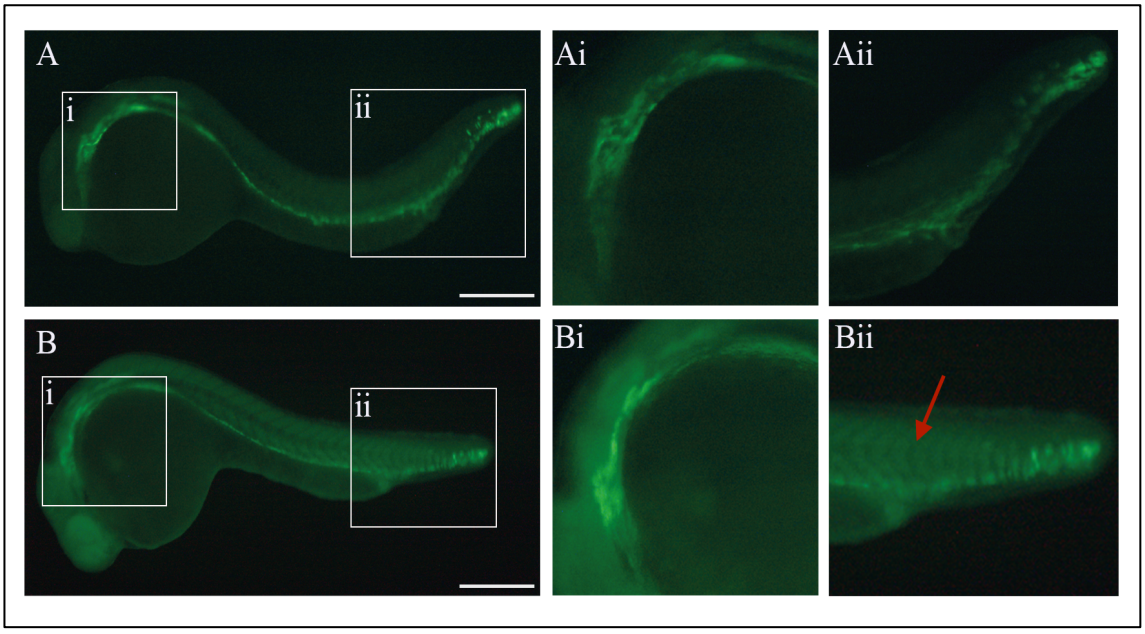


Figure 4.3 Non leaky and leaky *sox17:GFP* embryos at 24 hpf. (**A**) Lateral view (anterior to the left) of non leaky embryo at 24 hpf. Insets represent zoomed areas labelled accordingly. (**B**) As (A) except showing a leaky embryo. Red arrow: GFP ectopic expression in somites, however leaky embryos showed no apparent phenotype or delay in development. Scale bar represents 100 µm.

In any biological system expressing a fluorescence protein, some variation in intensity of the fluorophore is expected (Eijlander and Kuipers, 2013) and in some systems leaky signals have been linked to the intrinsic auto-fluorescent nature of cellular components, for example molecules such as NADH, FADH or lipofuscin (Andersson et al., 1998). To assess the

presence of a fully folded and active GFP and to allow effective discrimination of GFP signal from cellular autofluorescence, I performed anti-GFP immunohistochemistry (IHC), analysed the spatial domains of expression and compared, semi-quantitatively, the levels of GFP protein at 24 hpf (Figure 4.4).

The GFP expression domain, as indicated by blue stain, was observed to be diffuse in leaky embryos compared to non leaky, with expression seen to occur outside of the endodermal lineages, for example in the somites (Figure 4.4B,Bi). Furthermore, quantification of GFP (pixel density) using an automatized protocol (FIJI/ImageJ), confirmed that GFP expression in leaky embryos was significantly higher than in non leaky embryos (Figure 4.4C).

It is also possible to use similar automatized protocols (FIJI/ImageJ) to quantify the fluorescence emitted by leaky and non leaky embryos. I therefore quantified the GFP (green pixel intensity) in images taken of live embryos (Figure 4.6). The results of this analysis corroborate the observations made in fixed embryos subject to IHC.

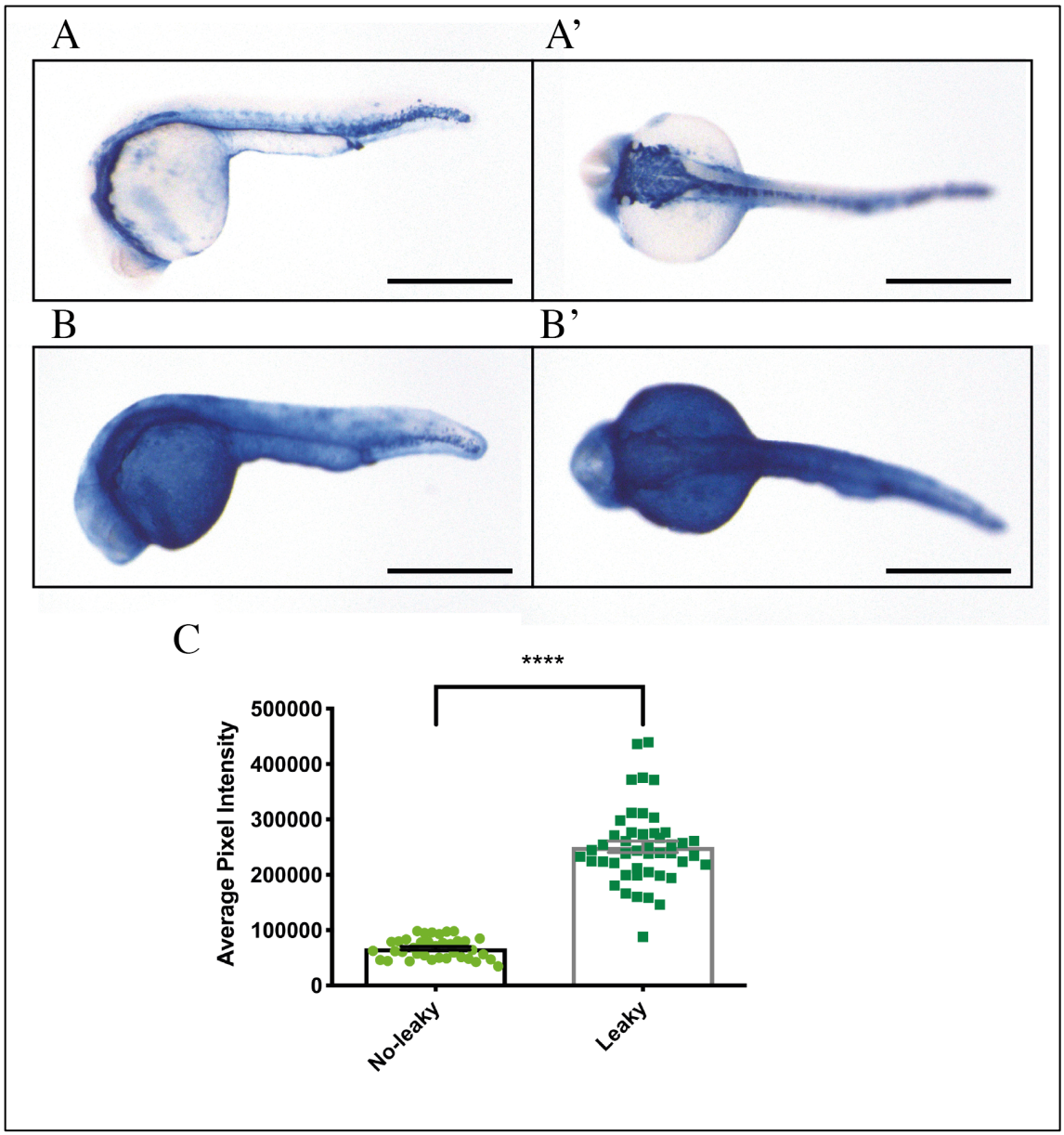


Figure 4.4 Anti-GFP immunohistochemistry on *sox17:GFP* embryos at 24 hpf. (**A**) and (**A'**) Lateral and dorsal views (anterior to the left) respectively of a 24 hpf non leaky embryo. **(B)** and **(B')** As for (A) and (A') except showing a leaky embryo. (**C**) Bar chart showing the average pixel intensity for the means of non leaky and leaky embryos from lateral view. Bars represent SEM. Statistical analysis performed using Student's t-test (two-tailed). ****$p \leq 0.001$. Scale bars represent 300 µm.

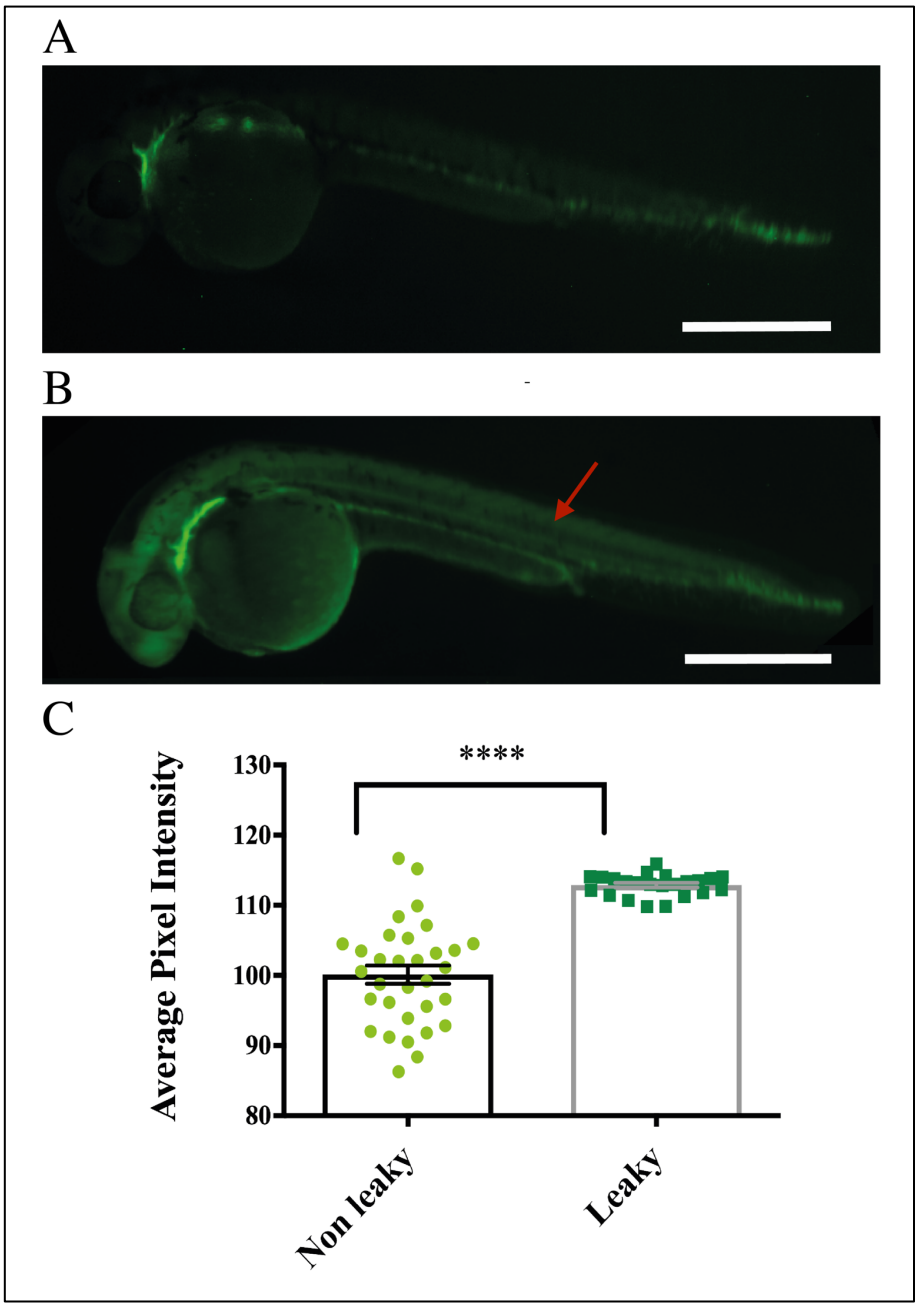


Figure 4.5 GFP quantification in non leaky and leaky embryos at 24 hpf. Quantification of GFP as green pixel intensity of (**A**) non leaky embryos at 24 hpf and (**B**) leaky embryos. Red arrow: GFP ectopic expression in somites. Lateral views with anterior to the left. (**C**) Bar chart showing the average pixel intensity for the means of non leaky and leaky embryos. Bars represent SEM. Statistical analysis performed using Student's t-test (two-tailed). ****$p \leq 0.001$. Scale bars represent 300 μm.

Taken together, the combination of IHC and fluorescent imaging supported the idea that natural autofluorescence was not interfering in image acquisition, rather, the data suggest the *sox17* promoter in leaky embryos was not tightly regulated, leading to *gfp* expression in non endodermal tissues such as the somites (Figure 4.3B and Figure 4.5B).

Recent studies have revealed crucial roles for maternal products for correct development of the germ layers, I then asked whether the leakiness trait I had observed was linked to maternal transmission and maternal regulation and whether the same females were consistently producing leaky embryos. I used the same female fish multiple times and expected that if the trait were maternally transmitted, all the progeny for each spawning should show leaky GFP expression. Remarkably, only 3% to 13% of embryos from each batch were leaky and some spawning combinations showed no leakiness at all (Figure 4.6A-D), causing me to reject the hypothesis that acquisition of the leakiness trait was linked to maternal transmission.

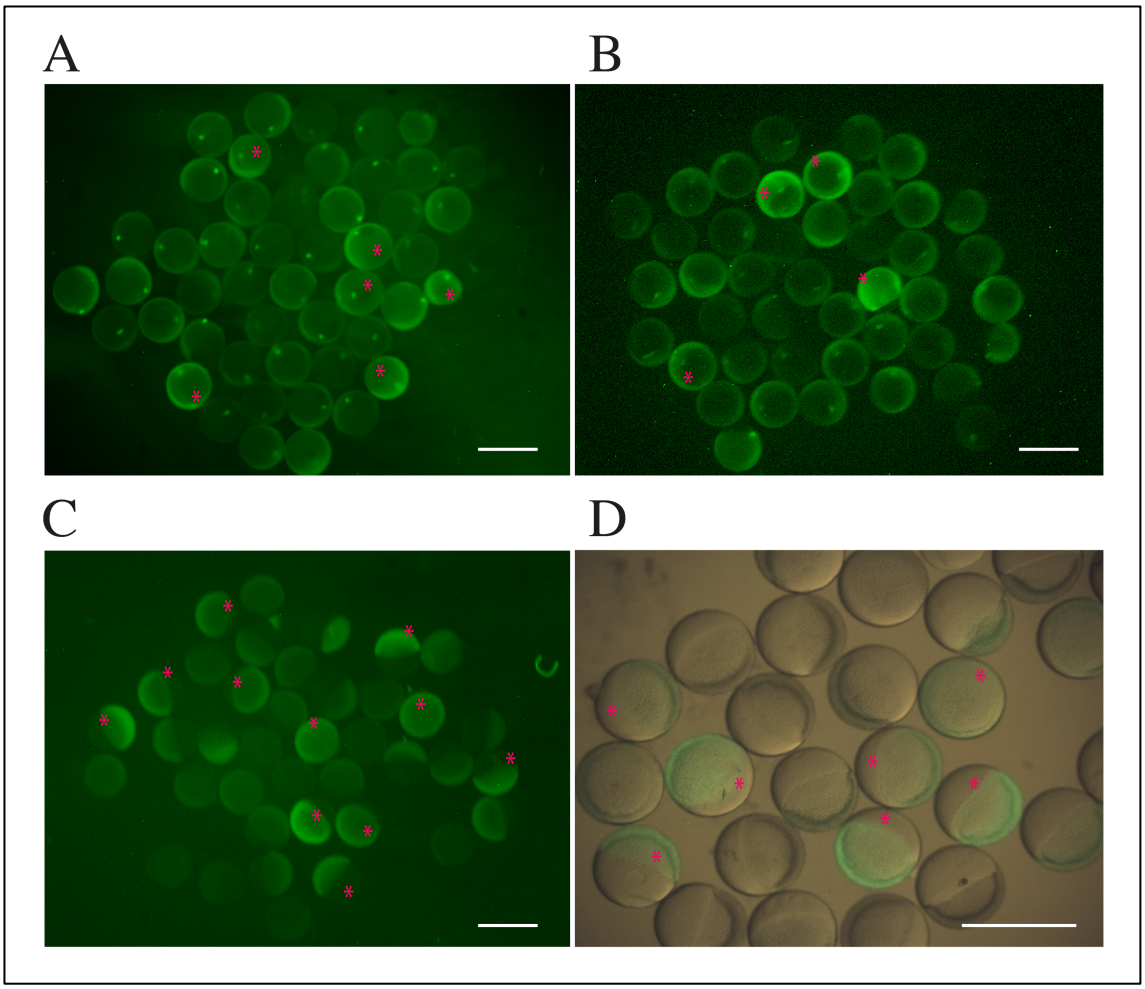


Figure 4.6 Proportion of leaky vs non leaky embryos from controlled *sox17:GFP* crosses. (**A**), (**B**) and (**C**) The proportion of inheritance of the leakiness trait varied from batch to batch and was therefore not maternally linked; three different batches are shown. Red asterisks indicate leaky embryos. (**D**) Overlay image showing batch variation under both brightfield and fluorescent microscopy. Scale bars represent 100 μm.

4.3 *gfp* mRNA transcript quantification

As immunochemistry showed that fluorescent protein distribution was significantly more concentrated and widespread in leaky embryos, I questioned whether the same pattern was detectable at the transcript level by RT-qPCR as well. Leaky and non leaky embryos were separated according to visual distribution of GFP as described above and I assessed the temporal dynamics of *gfp* expression at three key time points during development: the beginning of gastrulation (5.25 hpf) when *sox17* is expressed in the dorsal margin region, mid-gastrulation (7.00 hpf) and the end of gastrulation (9.00 hpf) when the level of *sox17* expression is at its peak. In addition, I also measured *gfp* expression at 24 hpf to assess whether any initial changes in transcripts number were maintained throughout somitogenesis. I decided to consider three independent biological batches, each being comprised of three individual embryos, to best represent the biological variability of the system (9 total embryos, see Material and Method, Figure 2.1 for more details). This decision was supported by the variability in pixel quantification in both fluorescent and IHC images. The quantity of *gfp* transcripts was determined using a standard curve, and the mean of non leaky batch one was used as the calibrator sample. *gfp* expression in all other samples was then expressed as a percentage, relative to the non leaky calibrator sample (see Material and Methods and paragraph 4.4 for more details). The leaky embryos expressed *gfp* at a significantly higher level than non leaky embryos at 5.25 hpf (1.4 fold increase) (Figure 4.7A), and this trend continued to be observed at 7.00 hpf (2 fold) (Figure 4.8B) and persisted throughout the end of gastrulation (2.8 fold) (Figure 4.8C**)**. These data suggested either instability of the regulatory system of the upstream region of the *sox17:GFP* construct (for example a mutation in the 5 kb regulatory sequence of the transgene) or an alternative malfunction/mechanism in the regulatory system (misregulation of temporal properties of TF activity). At the 7.00 hpf timepoint, although the trend towards higher *gfp* transcript percentage continued to be observed, due to higher variability in *gfp* values (higher SEM), the difference between batches 2 and 3 leaky and non leaky embryos did not reach statistical significance ($p = 0.06$). Expression of *gfp* was markedly lower in non leaky embryos compared to leaky embryos at all time points during gastrulation, but by 24 hpf no difference in *gfp* expression was observed (Figure 4.7D). These RT-qPCR results during gastrulation substantiated the observations I made using IHC approaches; I observed a high level of GFP positive cells in leaky embryos during gastrulation compared to non leaky (Figure 4.7). In contrast however, I also observed significant differences in the proportion of GFP positive cells at 24 hpf between leaky and non leaky embryos with IHC and

fluorescence microscopy (Figure 4.5C) but saw no difference at the transcript level (Figure 4.8D). There are many processes that occur between transcription and translation and as mentioned above, protein stability could be a big factor in explaining the discrepancy between mRNA and protein levels. Crucially, the apparent absence of *gfp* misregulation during somitogenesis may explain why no visible phenotype was observed in leaky embryos at 24 hpf (Figure 4.4, 4.5 and 4.6).

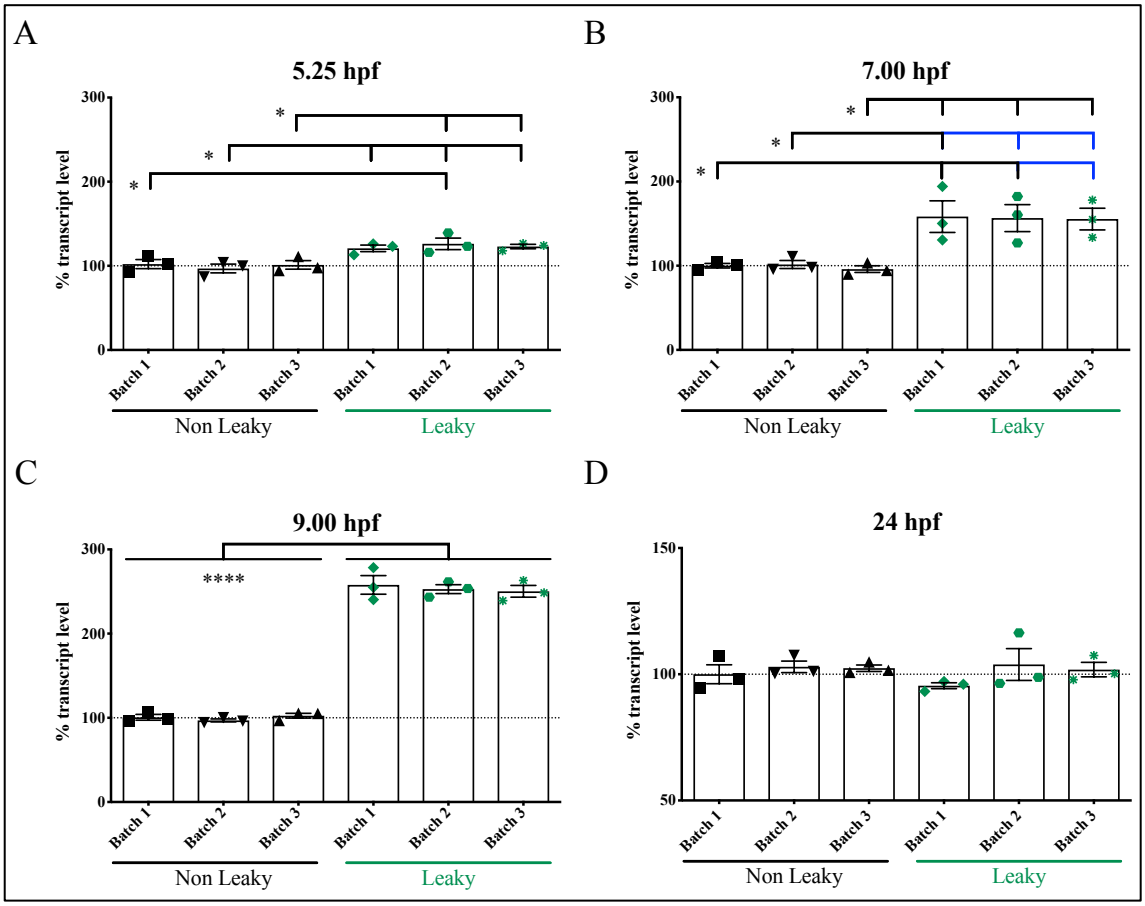


Figure 4.7 *gfp* transcript levels in non leaky vs leaky embryos. RT-qPCR analysis of *gfp* mRNA levels in non leaky and leaky embryos at (**A**) 5.25 hpf; (**B**) 7.00 hpf; (**C**) 9.00 hpf and (**D**) 24 hpf. (mean ± SEM, n = 3). *$p \leq 0.05$, ****$p \leq 0.001$, one-way ANOVA with Tukey's post-hoc test. Differences among batches observed at a $p = 0.06$ are reported in blue. By 7.00 hpf all three leaky embryos batches were significant upregulated and this pattern persisted at 9.00 hpf. Each batch consists of three individual embryos; each point represents an embryo and is the mean of two technical replicates.

4.4 Comparative expression of genes in leaky and non leaky embryos

As a consequence of the IHC/fluorescence imaging results and the *gfp* RT-qPCR, I questioned whether the leakiness of the *sox17:GFP* transgene could also lead to dysfunctional regulation of known genes that are active during gastrulation. I collected temporal expression data from ZFIN (Ruzicka et al., 2015) and merged them together with high-resolution mRNA expression time course data (White et al., 2017). From this, I selected candidate genes whose expression was tightly associated with endodermal, mesodermal and/or ectodermal fate and where expression lasted throughout gastrulation. A subset of these genes was also chosen as their expression persisted until at least 24 hpf (Figure 4.8). *nanog, sox19b* and *pou5f* genes were selected as pluripotency markers as reported in (Lee et al., 2013). *sox32, sox17*, *gata5* and *foxa2* were chosen to evaluate endodermal cells as they are well characterised markers (Alexander et al., 1999; Reiter et al., 1999; Sakaguchi et al., 2001). *mixl1* and *crcx4* are known markers of mesendodermal cells (Kikuchi et al., 2000; Mizoguchi et al., 2008) and *tbxta, myf5, vox,, tbx16, tbx24,* and *bmp4* are markers of different mesodermal derivates (Chocron et al., 2007; Imai et al., 2001; Jahangiri et al., 2012; Morley et al., 2009; Pownall et al., 2002; Stickney et al., 2007). Neuroectoderm and ectodermal lineages were assessed using *tfap2a, foxi1, irx7,* and *otx2* respectively (Knight et al., 2003; Koshida et al., 1998; Li and Cornell, 2007; Solomon et al., 2003).

During development, the temporal dynamics of gene expression are highly dynamic, and as germ layer precursor cells progressively separate into different spatial domains, the same genes are reused in different specification programs and at different stages of development, therefore expression of the same gene can be observed in multiple anatomical structures. The gene classification in Figure 4.8 is based on the main spatial expression domain of the gene during gastrulation, for example *gata5* is expressed in endodermal cells during gastrulation and later on at 24 hpf in the heart (Reiter et al., 1999; Reiter et al., 2001).

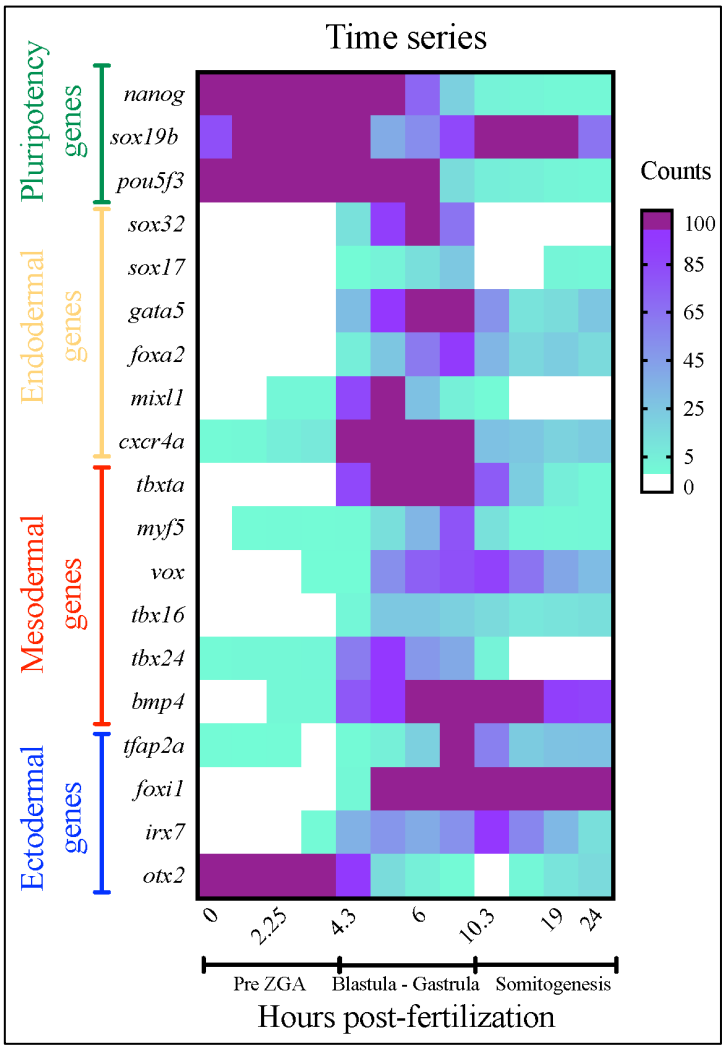


Figure 4.8 Time series heatmap. Expression of key genes involved in endoderm (yellow), mesoderm (red) and ectoderm (blue) specification during the first 24 hours of zebrafish development are shown. Three pluripotency markers are also included (green). Gene are categorised according to their main spatial expression domain during gastrulation. Data obtained from ZFIN and adapted from White et al., 2017.

In the first instance, I examined the expression of these gene sets in zebrafish by RT-qPCR. In order to compare gene expression profiles from leaky, non leaky and WT embryos I opted to use the standard curve method with the above panel of 12 genes of interest. I included a calibrator gene (*gfp*) and a well characterised reference gene, *elf2,* whose expression is stable across the above time series (Tang et al., 2007) and across the three conditions, leaky, non leaky and WT. In order to directly compare the RT-qPCR results, all gene expression data were normalised to the average of *elf2* expression.

In order to accurately use the standard curve-based transcript quantification method, cDNA template expressing the particular gene of interest in high abundance was used to generate a standard curve across five 5-fold serial dilutions. The GFP calibration curve was generated via

the same methodology using plasmid DNA. I plotted Ct value against dilution factor in a base-10 semi-logarithmic graph, fitting a line to the linear portion of the amplification curves. I confirmed that the correlation coefficient (R^2) for the line was > 0.98 and that all primers showed an efficiency of 100 ± 10%. These standard curves for each gene were then used to extrapolate the relative expression level for the same gene of interest in the different experimental samples. A standard curve for the reference gene (*elf2*) was produced separately and the relative quantification result for the gene of interest was normalised to that of the reference gene in the same sample. Normalised values were then compared to the WT samples and expressed as a % of WT transcript level.

To obtain consistent and accurate results which truly reflect mRNA expression levels in developing embryos, good controls are crucial for RT-qPCR (Van Peer et al., 2012). Therefore, in addition to evaluating the efficiency, correlation coefficient, precision, and sensitivity of the RT-qPCR, I also considered RNA input quality, DNA contamination and included no template and noRT controls. Assessing all these factors together allowed for a rigorous evaluation and comparison of variation in gene expression between the three conditions (WT, non leaky and leaky).

In terms of efficiency, I considered an acceptable rate to be between 95 and 110%; the absolute Ct value comparison is only meaningful when comparing experiments in which primer efficiency has been tested and is consistent. For the correlation coefficient, an R^2 value > 0.98 provided confidence in correlating Ct values with transcript abundance. A standard deviation of ≤ 0.4 among technical replicates confirmed precision of the method. In addition, the melt curve analysis needed to show a clean single peak, confirming that only the target species was being amplified. These stringent parameters allowed me to be sure that my primers were efficient and thus resulted in a proportional dose-response curve enabling me to determine the relative amount of gene expressed in experimental samples. For example, the regression equation for *sox17* was $y = -3.273 \log(x) + 32.694$ ($R^2 = 0.99$, $n = 3$); where a slope of –3.3 reflects a primers efficiency of 100%. In the same way, the amplification efficiency was 102.1% for *sox32* (Figure 4.9).

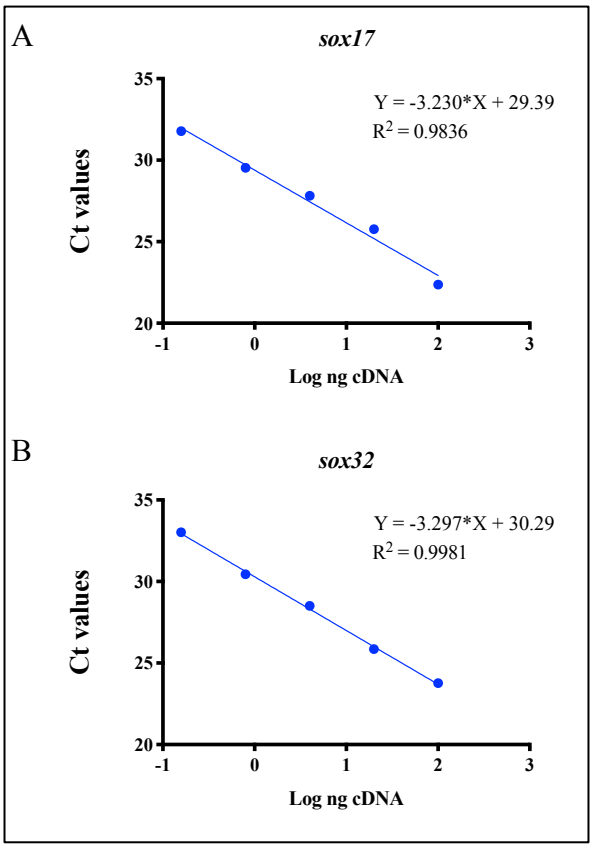


Figure 4.9 Standard curves for *sox17* (A) and *sox32* (B) RT-qPCR. All curves were based on serial dilution of cDNA. A slope of the curve around -3.3 means an efficiency of 100%. The R^2 of the curve should be > 0.98 to provide a good confidence within the correlation.

All primer efficiency and R^2 values for all RT-qPCR primers are shown in Table 4.1.

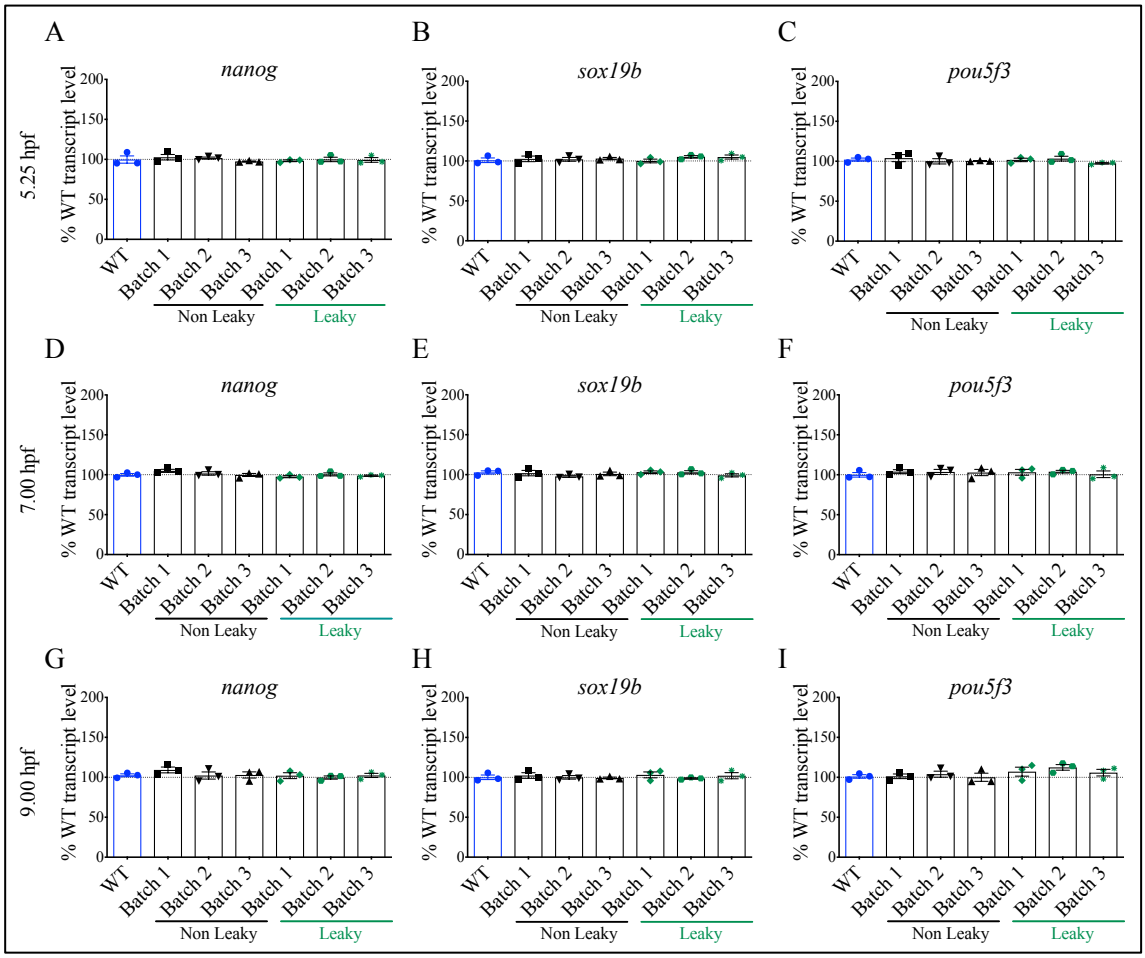
Table 4.1 Efficiency and R^2 values for all primers used to assess gene expression in WT, non leaky and leaky embryos.

	Efficiency	**R^2**
elf2	97.93	0.969
gfp	98.14	0.998
nanog	102.95	0.997
pau5f	99.74	0.995
sox2	99.93	0.986
sox32	102.10	0.998
sox17	101.88	0.989
gata5	101.41	0.997
mixl1	103.90	0.975
crcx4	95.88	0.999

foxa2	99.16	0.995
tbxta	103.41	0.997
myf5	94.60	0.989
vox	104.06	0.999
tbx16	102.49	0.988
tbx24	99.50	0.983
bmb4	103.28	0.993
tfap2b	96.59	0.956
otx2	98.16	0.992
foxi1	100.41	0.998
sox2	97.92	0.998
irx7	101.88	0.984

4.4.1 Expression of pluripotency markers in leaky and non leaky embryos

Lee et al. (2013) and Leichsenring et al. (2013) previously showed that the expression of three pluripotency TFs such as Nanog, Sox19b (an ortholog of Sox2 in the SoxB1 family) and Pou5f3 (also known as Oct4) are associated with the maternal to zygotic transition in zebrafish that occurs between 512-cell (2.75 hpf) and dome (4.3 hpf) stage and alteration in level of these genes are able to affect transcript levels post ZGA at 4.5 hpf and 5.3 hpf (Meier et al. 2017). I therefore first questioned whether the misregulation of *gfp* was affecting these known pluripotency genes in leaky embryos (Figure 4.10). No discernible difference was observed for any of the pluripotency genes tested in both non leaky and leaky embryos compared to WT at any developmental stage (5.25 hpf, 7.00 hpf, 9.00 hpf). The unchanged expression of pluripotency markers at all four time points suggested that initial cell lineage commitment was not affected, even in leaky embryos. These data suggest that control of *gfp* regulation is associated with other mechanisms and has no bearing on the expression of the early pluripotency markers tested.



4.10 Pluripotency markers in non leaky and leaky embryos vs WT. Data in graphs are represented as mean ± standard error of the mean (SEM) with each data point representing one embryo and is the mean of two technical replicates. One-way ANOVA with Tukey's post-hoc test was used to assess significant differences in the levels of 3 genes (*nanog, sox19b* and *pou5f3*), as labelled, associated with pluripotency in zebrafish. No difference was observed at all 3 time points.

4.4.2 Expression of endodermal markers in leaky and non leaky embryos

As *gfp* expression is under the control of the *sox17* promoter, I next questioned whether other endodermal gene (*sox32, sox17, gata5, mixl1, cxcr4* and *foxa2*) expression was affected. At 5.25 hpf, the expression of *cxcr4* ($p \leq 0.01$) and *mixl1* ($p \leq 0.05$) were found to be significantly misregulated in leaky embryos compared to both WT and non leaky embryos (Figure 4.11E, F). This downregulation was not observed at later stages of gastrulation (7.00 hpf onwards).

At 5.25 hpf a slight upregulation of *sox17* was notable (Figure 4.11A) and *sox32* expression was relatively more abundant in leaky embryos, but this trend was not significant at this earlier stage (Figure 4.11B). In contrast, by 7.00 hpf expression of both genes were significantly

upregulated in leaky embryos compared to non leaky and WT. (Figure 4.12A, B). This disparity between leaky and non leaky embryos continued to increase and by 9.00 hpf three times more *sox17* transcripts and four times more *sox32* transcripts were observed in leaky embryos compared to non leaky and WT (Figure 4.13A, B). I also observed higher levels of these transcripts using WISH, thereby ratifying these RT-qPCR data (Figure 4.28). These observed changes in gene expression all coincided with the upregulation of *gfp* expression (Figure 4.7).

In the literature model of the endoderm pathway, *sox32* is upstream of *sox17* (Alexander et al., 1999; Aoki et al., 2002; Dickmeis et al., 2001). My data suggested that in leaky embryos this linearity was somehow inverted/not respected: in WT embryos, *sox32* transcripts were present at 4.5 hpf, with *sox17* expression identified at around 5.00 hpf (but at a much lower level than *sox32*). Using absolute quantification methodology, my data showed that the level of *sox17* transcripts was actually higher than that of the *sox32* transcripts at the 5.25 hpf time point in leaky embryos. In addition, the expression of the pluripotency marker *pou5f*, a known co-factor of *sox32*, showed no discernible difference in expression in leaky embryos, despite the observed upregulation of *sox32* (Lunde et al., 2004; Perez-Camps et al., 2016). Even at mid-gastrula stage, when *sox32* levels became strongly upregulated in leaky embryos, *pouf53* expression is comparable to WT, suggesting that Sox32 is controlled in an independent manner and this raises the possibility of more 'fine-tuning' players in the endoderm cascade.

In contrast to *sox32*, the expression of the endodermal genes *gata5* and *foxa2* showed no change between leaky, non leaky and WT (Figures 4.11, 4.12, 4.13C, D). Both *gata5* and *foxa2* control *sox32* expression and indirectly, as a cascade, *sox17* (Aoki et al., 2002; Nelson et al., 2017). The results suggest that multiple endodermal regulatory mechanisms are present in zebrafish and that the TFs involved in lineage specification control target gene expression in a synergistic, non-linear manner.

Consistent with leaky *gfp* expression, all misexpressed endodermal genes showed normal gene expression by 24 hpf, and no phenotypic defects were observed (Figure. 4.14). This interesting observation may suggest the existence of a regulatory mechanism within the embryo that can compensate for early, aberrant gene expression.

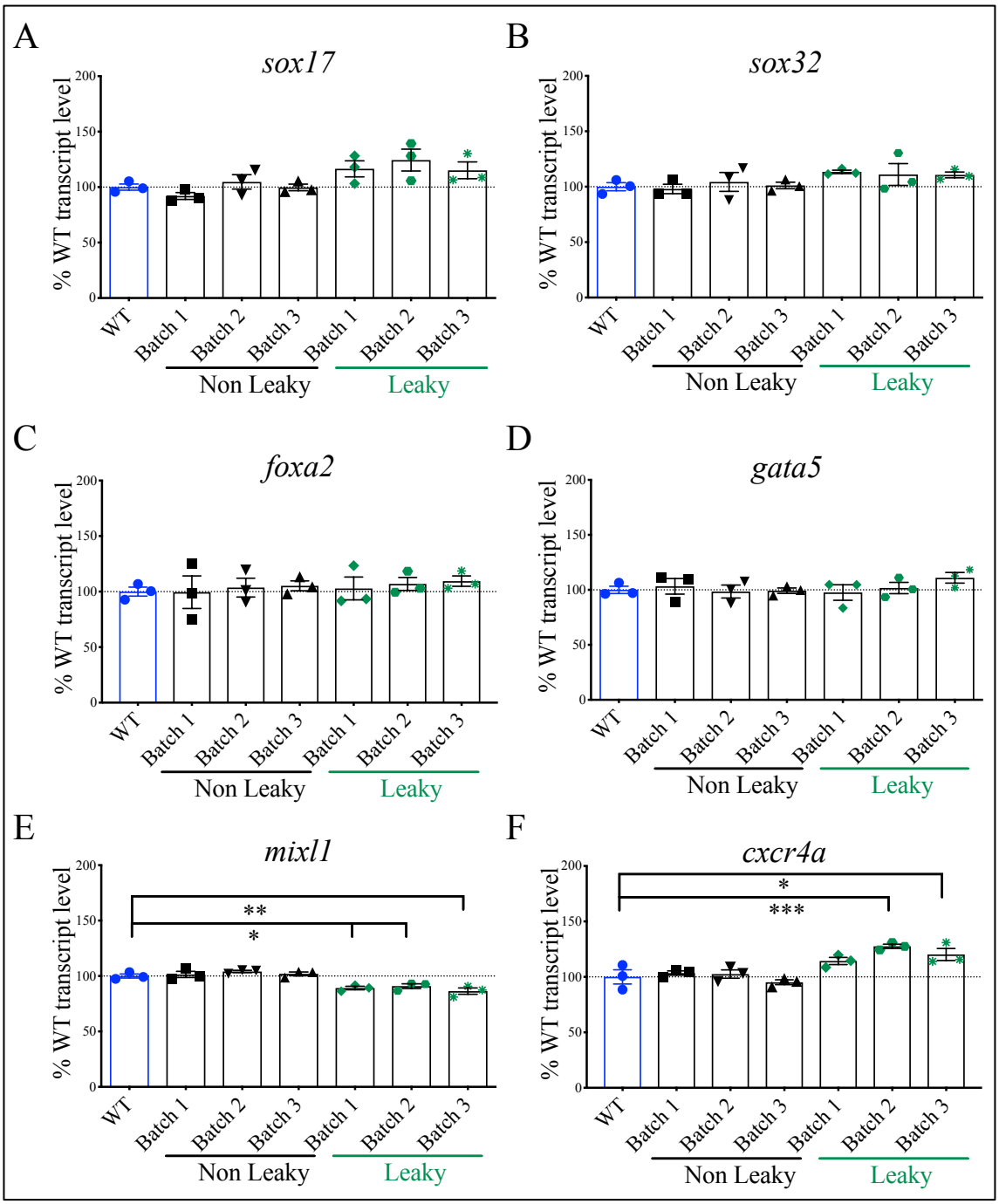


Figure 4.11 Expression of endodermal genes in leaky vs non leaky embryos at 5.25 hpf. (A) – (F)

Expression of endodermal genes, as labelled, as determined by RT-qPCR. The upregulation of *gfp* in leaky embryos coincided with a slight upregulation of *sox17* (A) and *cxcr4a* (F) and significant downregulation of other early mesendoderm marker *mixll* (E). Reference gene *elf2* expression was used for normalisation. The assessment was based on three independently derived batches; each batch consisted of three individual embryos; each point represents an embryo and is the mean of two technical replicates. Values are presented as % of WT expression ± SEM. Statistical analysis was performed using one-way ANOVA with Tukey's post-hoc test; *$p \leq 0.05$, **$p \leq 0.01$, ***$p \leq 0.001$.

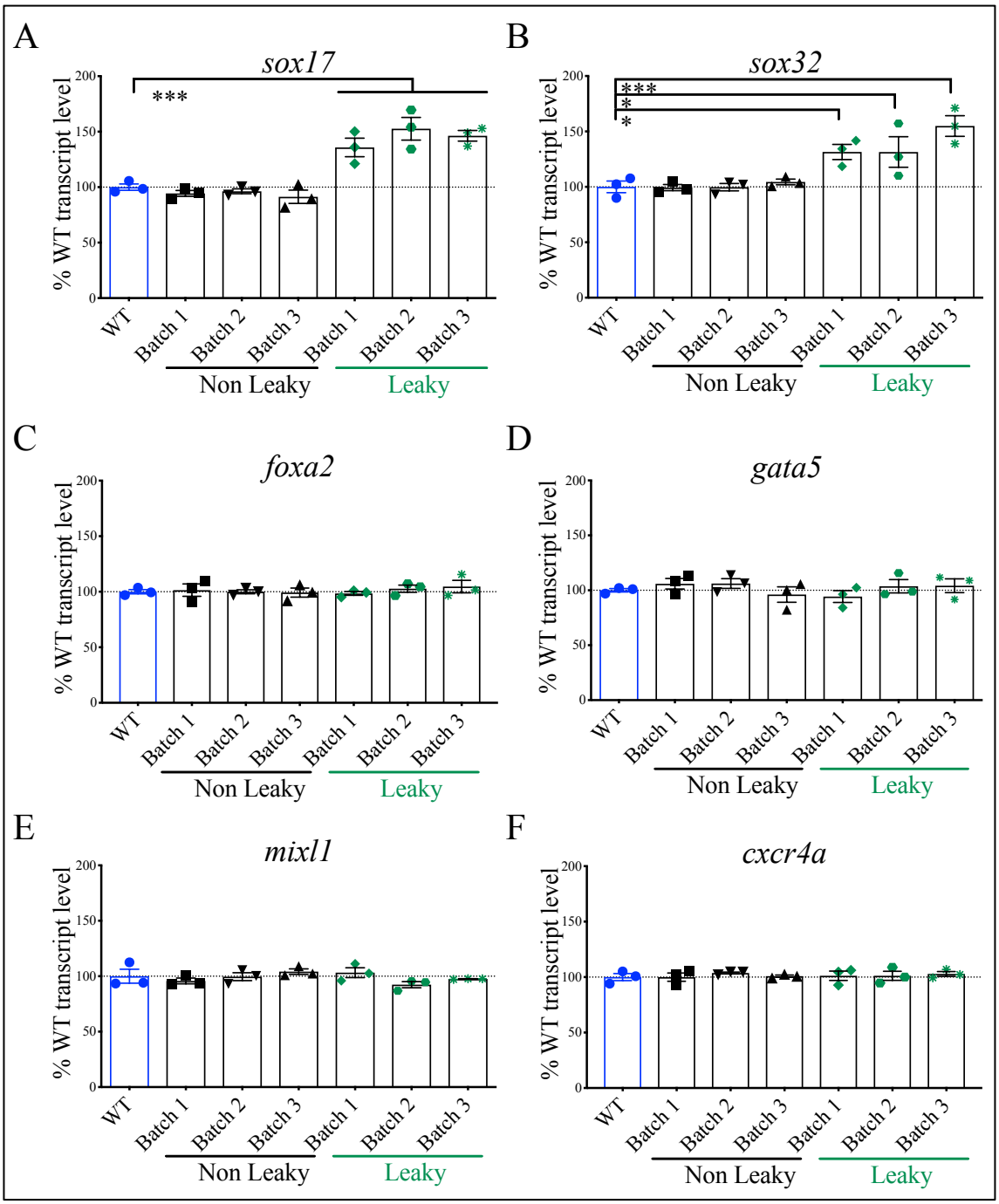


Figure 4.12 Expression of endodermal genes in leaky vs non leaky embryos at 7.00 hpf. (A) – (F)

Expression of endodermal genes, as labelled, as determined by RT-qPCR. The upregulation of *gfp* in leaky embryos coincides with a significant upregulation of *sox17* (A) and *sox32* (B). At this time point, no significant up or downregulation of other early markers was observed. Reference gene *elf2* expression was used for normalisation. The assessment was based on three independently derived batches; each batch consists of three individual embryos; each point represents an embryo and is the mean of two technical replicates. Values are presented as % of WT expression ± SEM. Statistical analysis was performed using one-way ANOVA with Tukey's post-hoc test; *$p \leq 0.05$, ***$p \leq 0.001$.

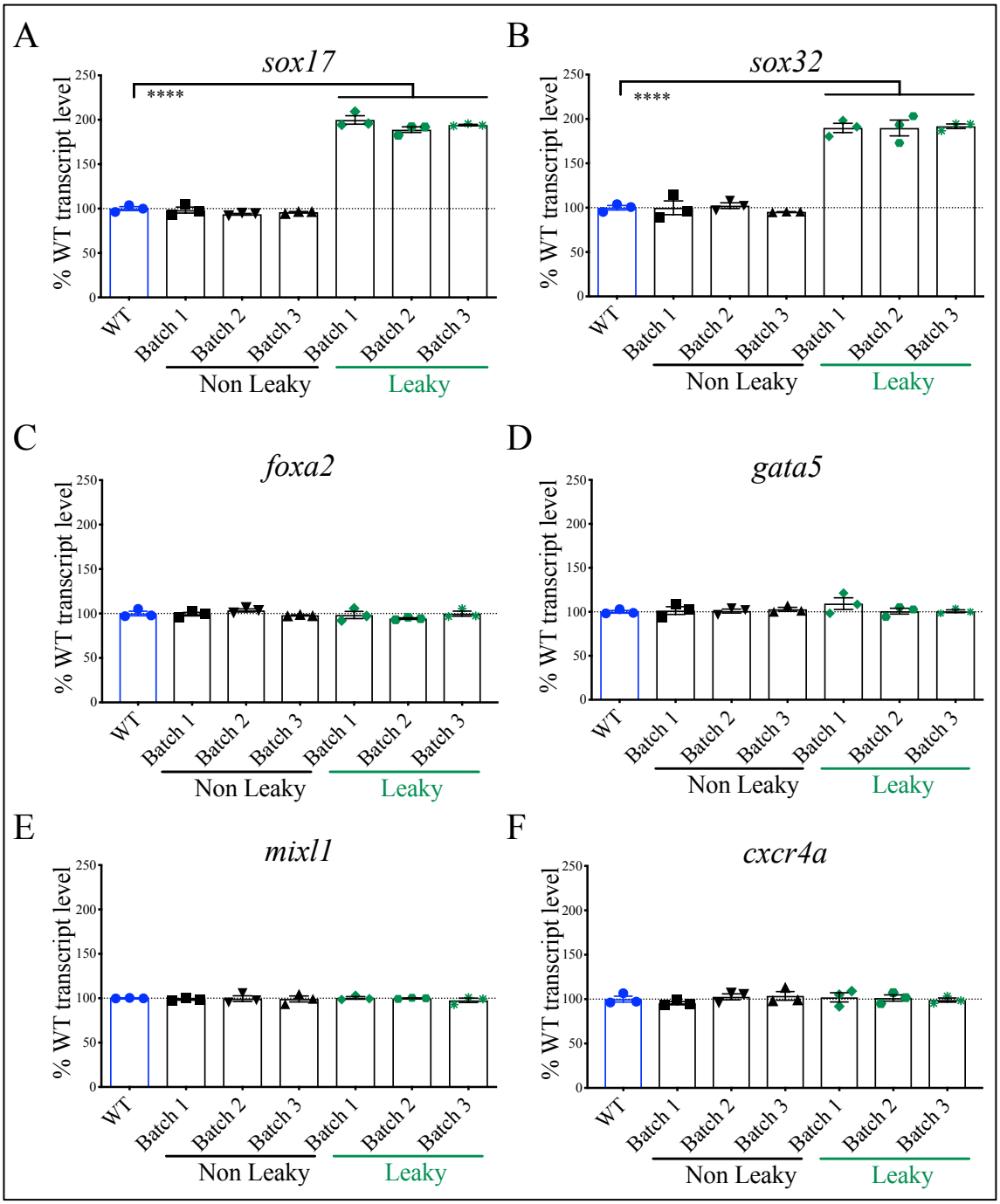


Figure 4.13 Expression of endodermal genes in leaky vs non leaky embryos at 9.00 hpf. (A) – (F)

Expression of endodermal genes, as labelled, as determined by RT-qPCR. The upregulation of *gfp* in leaky embryos coincides with a significant upregulation of *sox17* (A) and *sox32* (B). At this time point, no significant up or downregulation of other early markers was observed. Reference gene *elf2* expression was used for normalisation. The assessment was based on three independently derived batches; each batch consists of three individual embryos; each point represents an embryo and is the mean of two technical replicates. Values are presented as % of WT expression ±SEM. Statistical analysis was performed using one-way ANOVA with Tukey's post-hoc test; ****$p \leq 0.0001$.

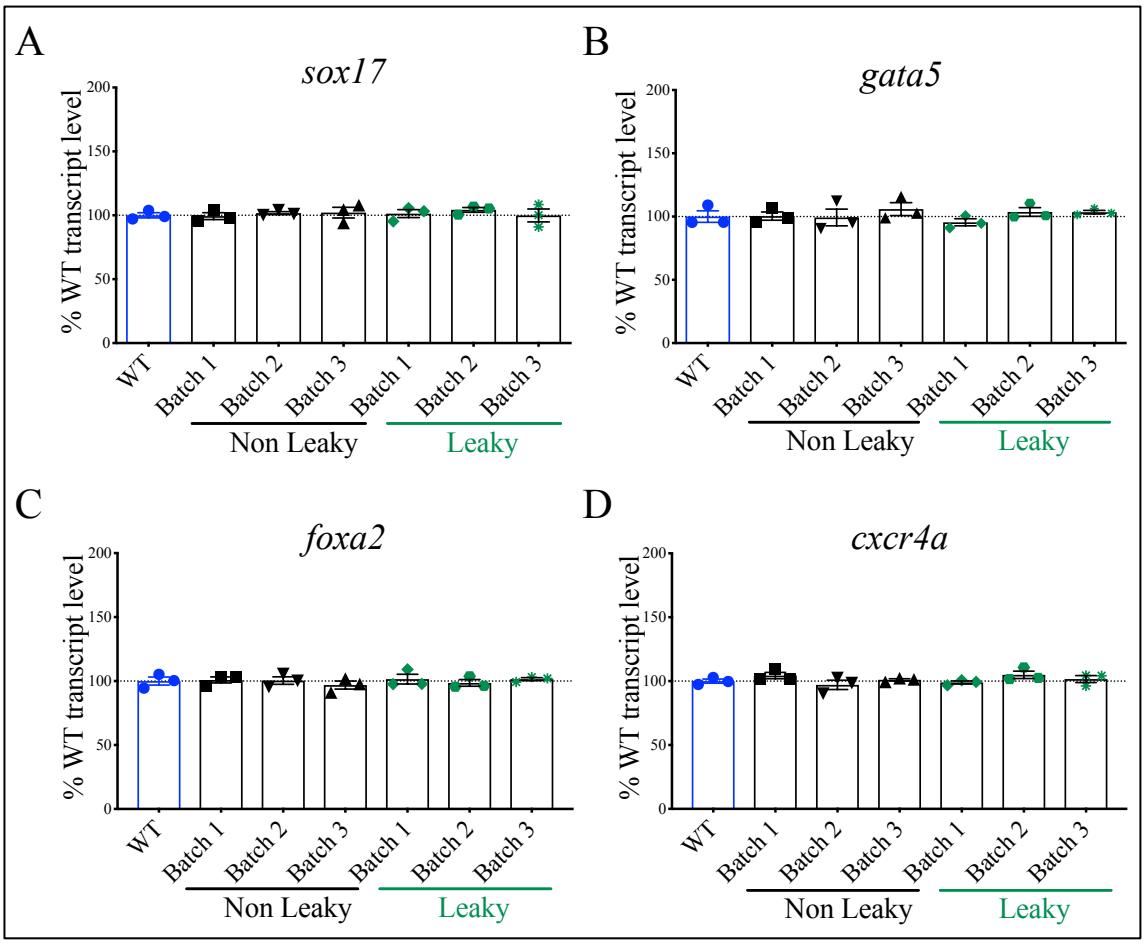


Figure 4.14 Expression of endodermal genes in leaky vs non leaky embryos at 24.00 hpf. (A) – (F) Expression of endodermal genes, as labelled, as determined by RT-qPCR. By 24 hpf, no significant up or downregulation of any marker tested was observed. Note that *sox32* and *mixl1* were no longer expressed at this time point. Reference gene *elf2* expression was used for normalisation. The assessment was based on three independently derived batches; each batch consists of three individual embryos; each point represents an embryo and is the mean of two technical replicates. Values are presented as % of WT expression ± SEM.

4.4.3 Expression of non endodermal markers in leaky and non leaky embryos

As increased *gfp* expression in leaky embryos coincided with a significant increase in *sox32* and *sox17*, I next asked whether markers of other lineages, i.e. mesoderm and ectoderm, were affected. It is known that differentiation towards a specific lineage often either results in, or is caused by, inhibition of an alternative lineage (Mizoguchi et al., 2006; Poulain et al., 2006; van Boxtel et al., 2018), and I therefore questioned whether the increase in endodermal gene expression observed in leaky embryos could affect the expression of other lineage specific genes. To address this question, I undertook RT-qPCR analysis as previously described on leaky, non leaky and WT embryos using markers which denote specification towards mesoderm and ectoderm (Figure 4.8).

4.4.3.1 Mesodermal markers

The mesodermal markers I selected for analysis are expressed in different mesodermal cell derivatives and cover axial mesoderm/notochord and paraxial mesoderm/future somites. At the initial 5.25 hpf timepoint, no significant difference was observed in expression of any of the mesodermal markers tested (Figure 4.15), however, a trend towards down-regulation of *tbxta* and *myf5* was apparent in leaky embryos. By the mid-gastrula stage (7.00 hpf), upregulation of *gfp* transcripts in leaky embryos coincided with significant down-regulation of the mesodermal marker *myf5* ($p \leq 0.0001$) (Figure 4.16). This downregulation continued to be observed at 9.00 hpf (Figure 4.17) and thus persisted throughout gastrulation. In contrast, no significant change was observed between leaky and non leaky embryos for all other mesodermal markers (*tbxta, vox, tbx24* and *tbx16*) at any time point; nevertheless, expression of *tbxta* showed a trend toward reduced expression and it is possible that higher variability in the non leaky biological replicates obscured this pattern (Figures 4.15, 4.16, 4.17). These data suggested that regulation of these other, unchanged, mesodermal genes was not affected in leaky embryos.

At the 24 hpf timepoint, no change was observed between leaky and non leaky embryos for any of the genes tested (Figure 4.18), including *myf5*, which was previously seen to be down-regulated throughout gastrulation. Indeed, the previous down-regulation of *myf5*, a muscle marker, did not lead to any apparent muscle-related phenotype during somitogenesis, possibly reflecting the fact that at 24 hpf, there was no difference in *myf5* transcript levels between leaky, non leaky and WT embryos (Figure 4.18B). This interesting phenomenon of gene expression stabilisation reiterated that seen for the misregulated endodermal genes, which were also seen to revert to WT expression levels by 24 hpf (Figure 4.14).

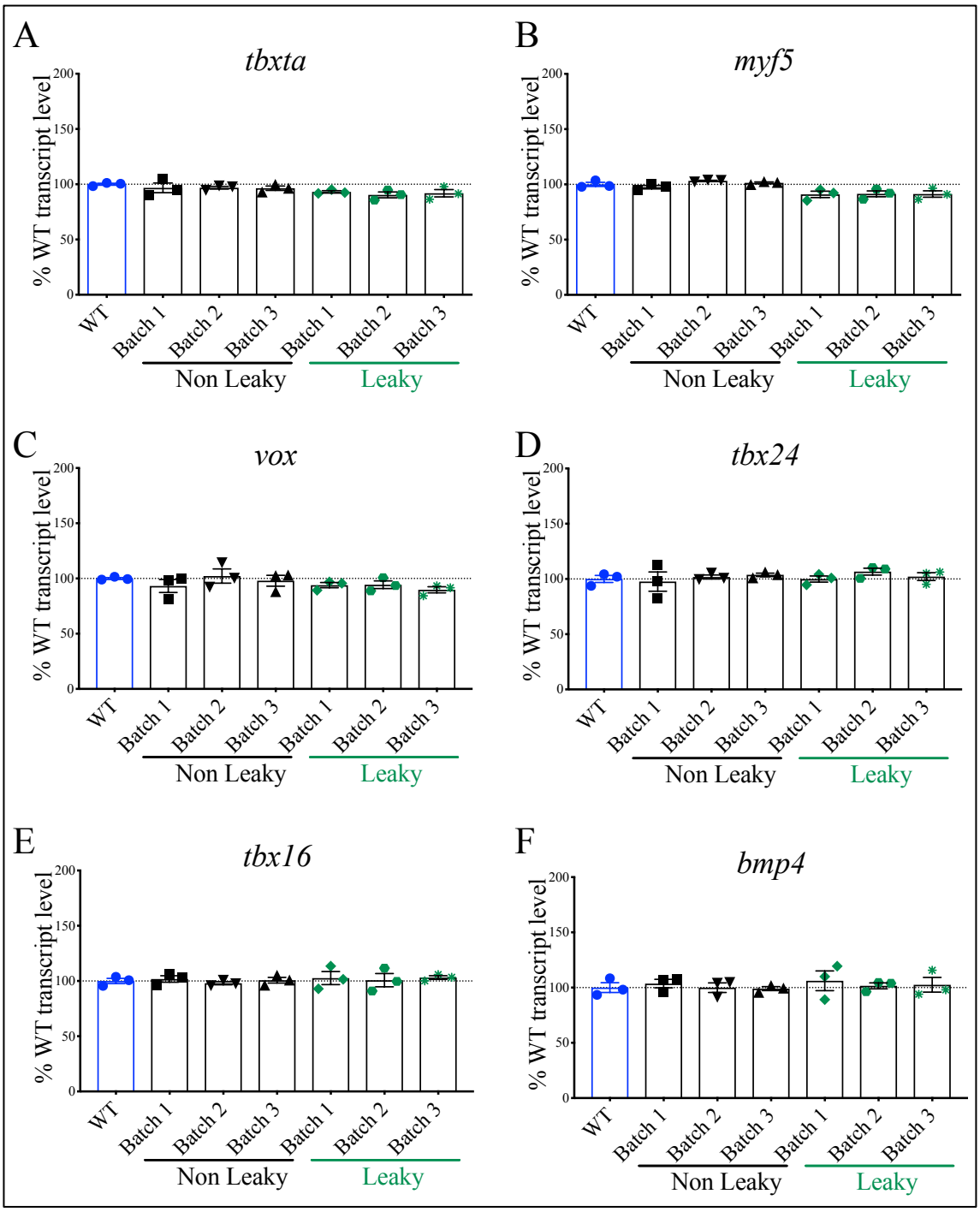


Figure 4.15 Expression of mesodermal genes in leaky vs non leaky embryos at 5.25 hpf. (A) – (F) Expression of mesodermal genes, as labelled, as determined by RT-qPCR. At this time point, no significant up or downregulation of mesodermal markers was observed. Reference gene *elf2* expression was used for normalisation. The assessment was based on three independently derived batches; each batch consists of three individual embryos; each point represents an embryo and is the mean of two technical replicates Values are presented as % of WT expression ± SEM.

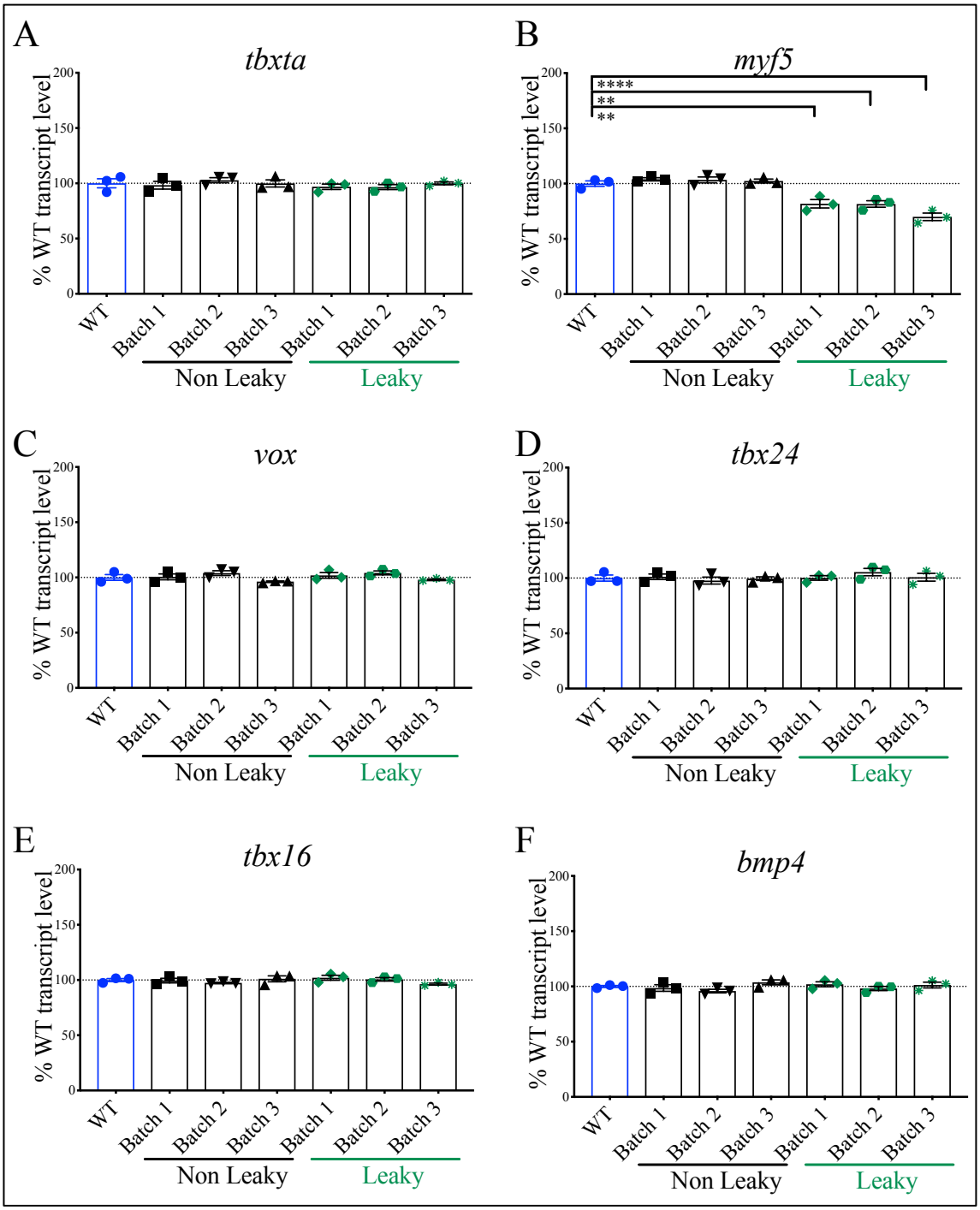


Figure 4.16 Expression of mesodermal genes in leaky vs non leaky embryos at 7.00 hpf. (A) – (F)

Expression of mesodermal genes, as labelled, as determined by RT-qPCR. The upregulation of *gfp* in leaky embryos coincided with a significant downregulation of *myf5* (B). Reference gene *elf2* expression was used for normalisation. The assessment was based on three independently derived batches; each batch consists of three individual embryos; each point represents an embryo and is the mean of two technical replicates. Values are presented as % of WT expression ± SEM. Statistical analysis was performed using one-way ANOVA with Tukey's post-hoc test; *$p \leq 0.05$, **$p \leq 0.01$, ****$p \leq 0.001$.

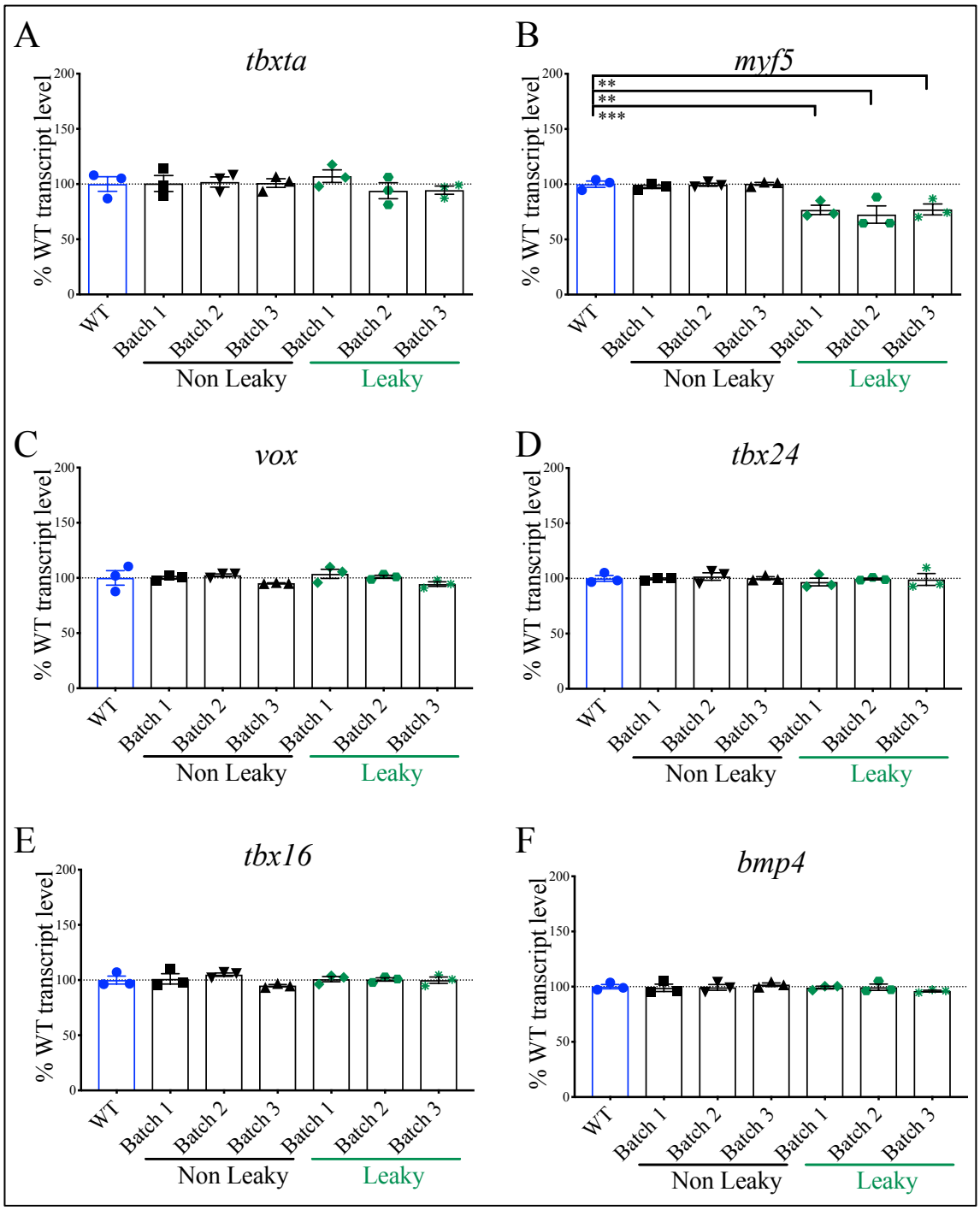


Figure 4.17 Expression of mesodermal genes in leaky vs non leaky embryos at 9.00 hpf. (A) – (F) Expression of mesodermal genes, as labelled, as determined by RT-qPCR. The upregulation of *gfp* in leaky embryos continued to coincide with the significant downregulation of *myf5* (B). Reference gene *elf2* expression was used for normalisation. The assessment was based on three independently derived batches; each batch consists of three individual embryos; each point represents an embryo and is the mean of two technical replicates). Values are presented as % of WT expression ± SEM. Statistical analysis was performed using one-way ANOVA with Tukey's post-hoc test; ** $p \leq 0.01$.

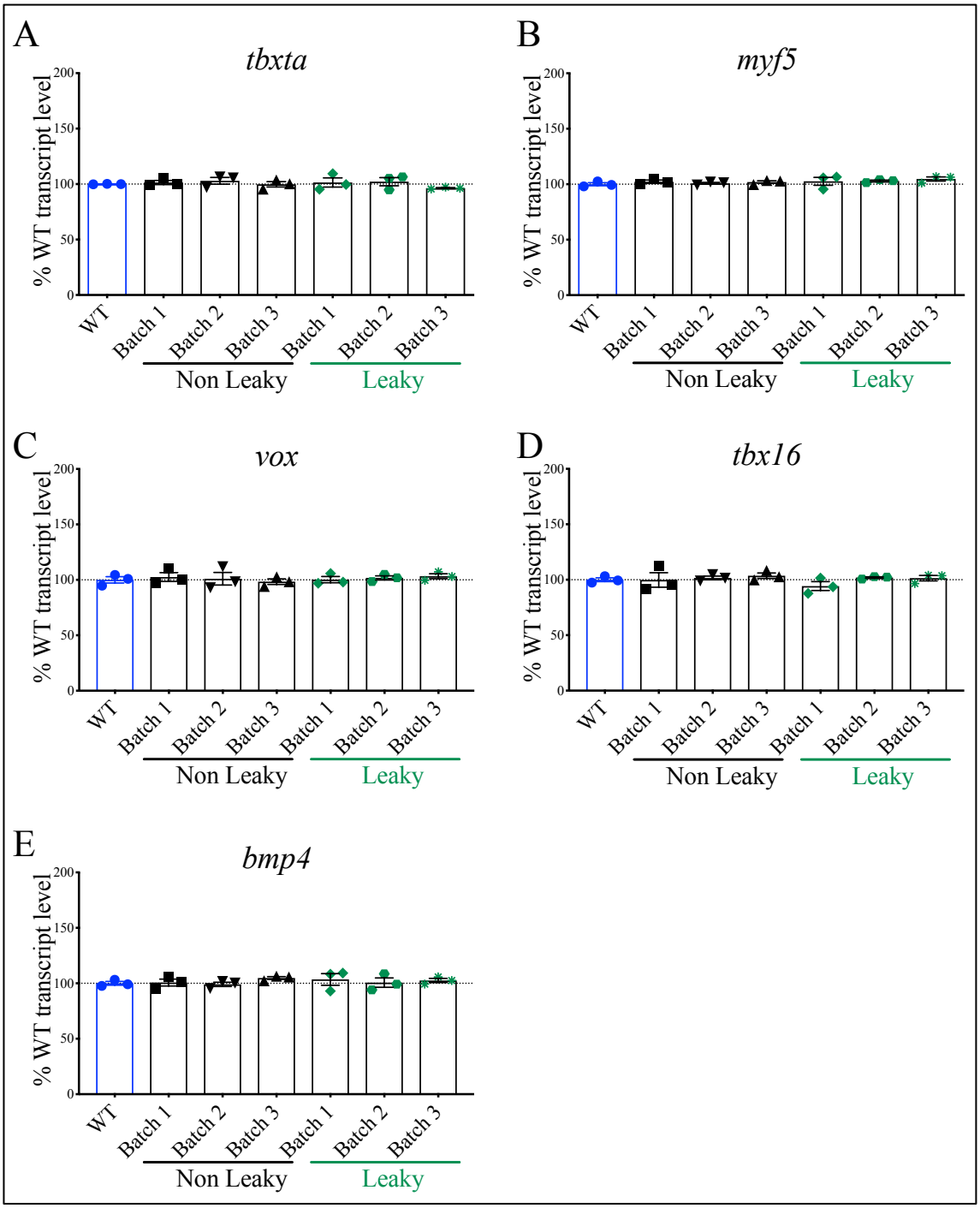


Figure 4.18 Expression of mesodermal genes in leaky vs non leaky embryos at 24.00 hpf. (A) – (D) By 24 hpf, no significant up or down-regulation of any marker tested is observed. Note that *tbx24* is no longer expressed at this time point). Reference gene *elf2* expression was used for normalisation. The assessment was based on three independently derived batches; each batch consists of three individual embryos; each point represents an embryo and is the mean of two technical replicates (n = 3, embryos = 9). Values are presented as % of WT expression ±SEM.

4.4.3.2 Ectodermal markers

In contrast to the observations made with endodermal and mesodermal gene expression, I observed no variations in expression of the ectoderm markers *tfap2, foxi1, irx7,* and *otx2* relative to WT and non leaky embryos at all time points tested (Figures 4.19-4.22), suggesting that whatever caused *gfp* to be misregulated in leaky embryos did not affect ectodermal gene expression. Pluripotent cells in the embryo initially differentiate into either ectoderm or mesendoderm, with the latter later segregating into mesoderm and endoderm (Hashimshony et al., 2015). This early lineage specification, and the closer temporal relationship between endoderm and mesoderm, may help to explain why misregulation of the *sox17* promoter leads to changes in mesodermal, but not ectodermal, gene expression. The speculation was then assessed capturing the whole transcriptomic changes in leaky embryos and the differential expression analysis showed ectodermal genes being unaffecting by the 'leaky' condition; see Chapter 5.

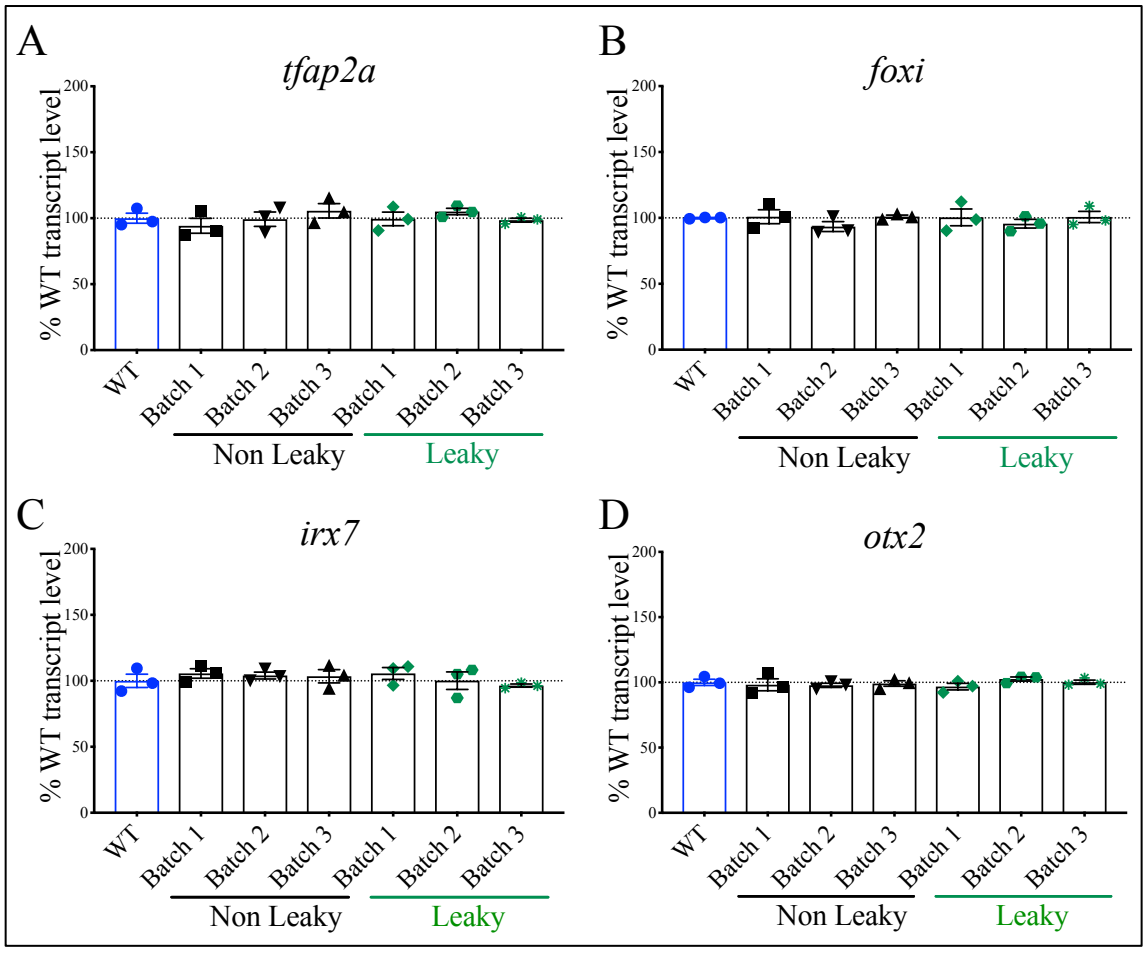


Figure 4.19 Expression of ectodermal genes in leaky vs non leaky embryos at 5.25 hpf. (A) – (D) No significant up or downregulation of any marker tested was observed. Reference gene *elf2* expression was used

for normalisation. The assessment was based on three independently derived batches; each batch consists of three individual embryos; each point represents an embryo and is the mean of two technical replicates. Values are presented as % of WT expression ± SEM.

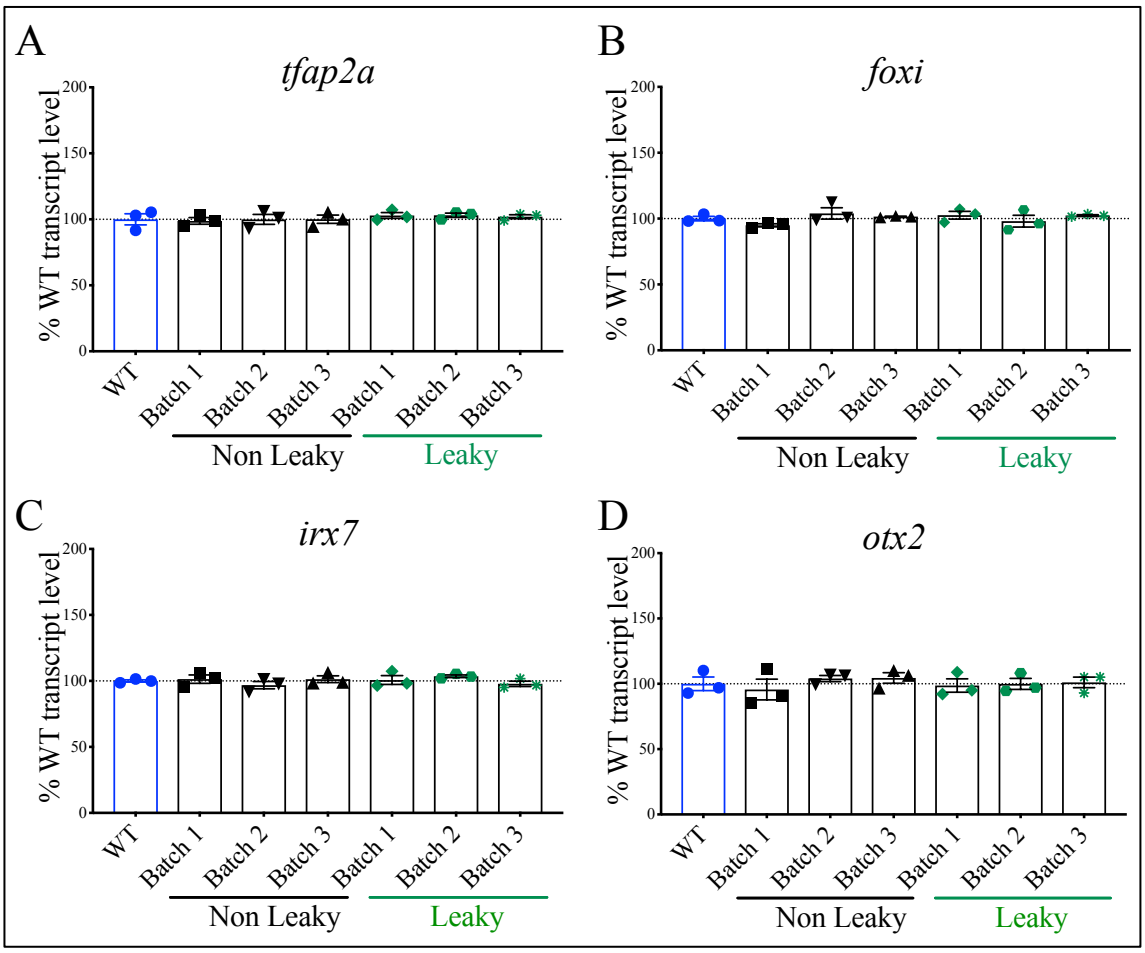


Figure 4.20 Expression of ectodermal genes in leaky vs non leaky embryos at 7.00 hpf. (A) – (D) No significant up or downregulation of any marker tested was observed. Reference gene *elf2* expression was used for normalisation. The assessment was based on three independently derived batches; each batch consists of three individual embryos; each point represents an embryo and is the mean of two technical replicates. Values are presented as % of WT expression ± SEM.

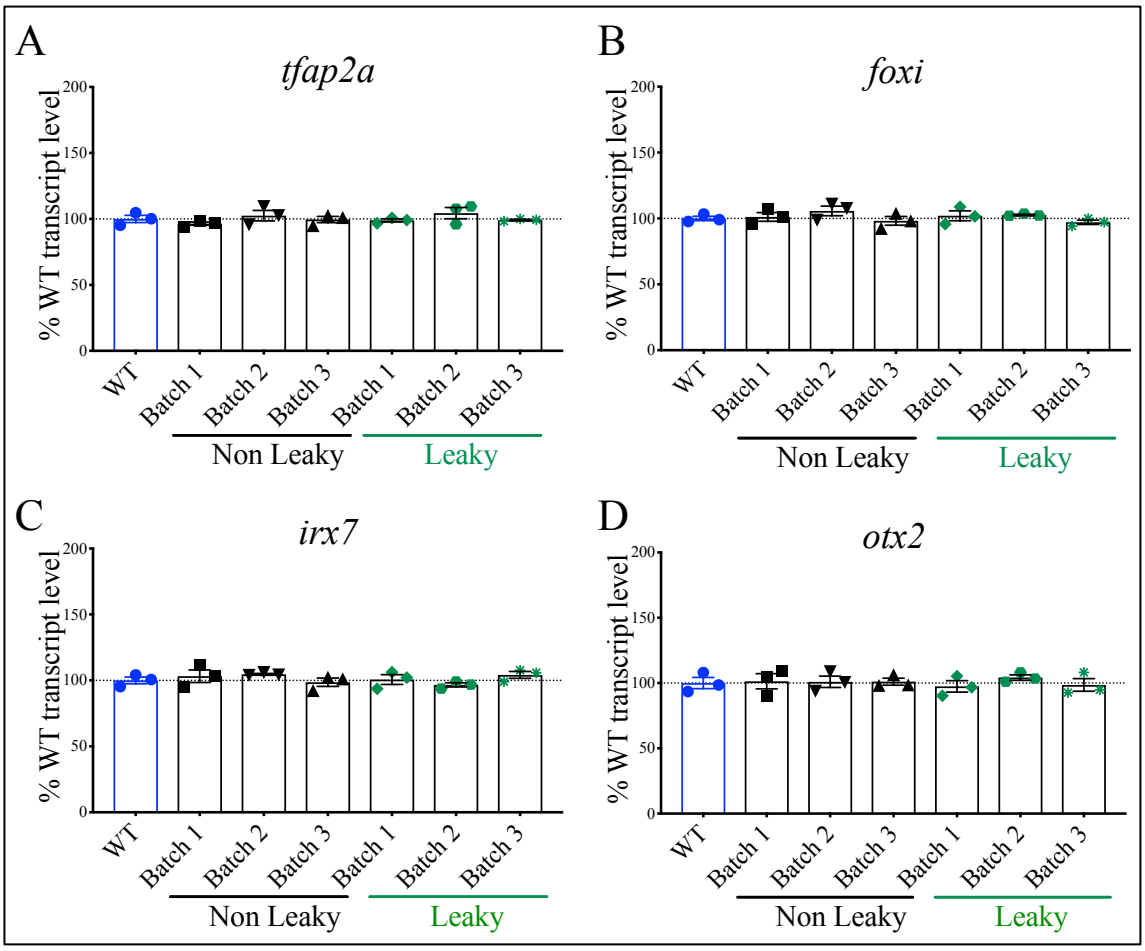


Figure 4.21 Expression of ectodermal genes in leaky vs non leaky embryos at 9.00 hpf. (A) – (D) No significant up or downregulation of any marker tested was observed. Reference gene *elf2* expression was used for normalisation. The assessment was based on three independently derived batches; each batch consists of three individual embryos; each point represents an embryo and is the mean of two technical replicates. Values are presented as % of WT expression ± SEM.

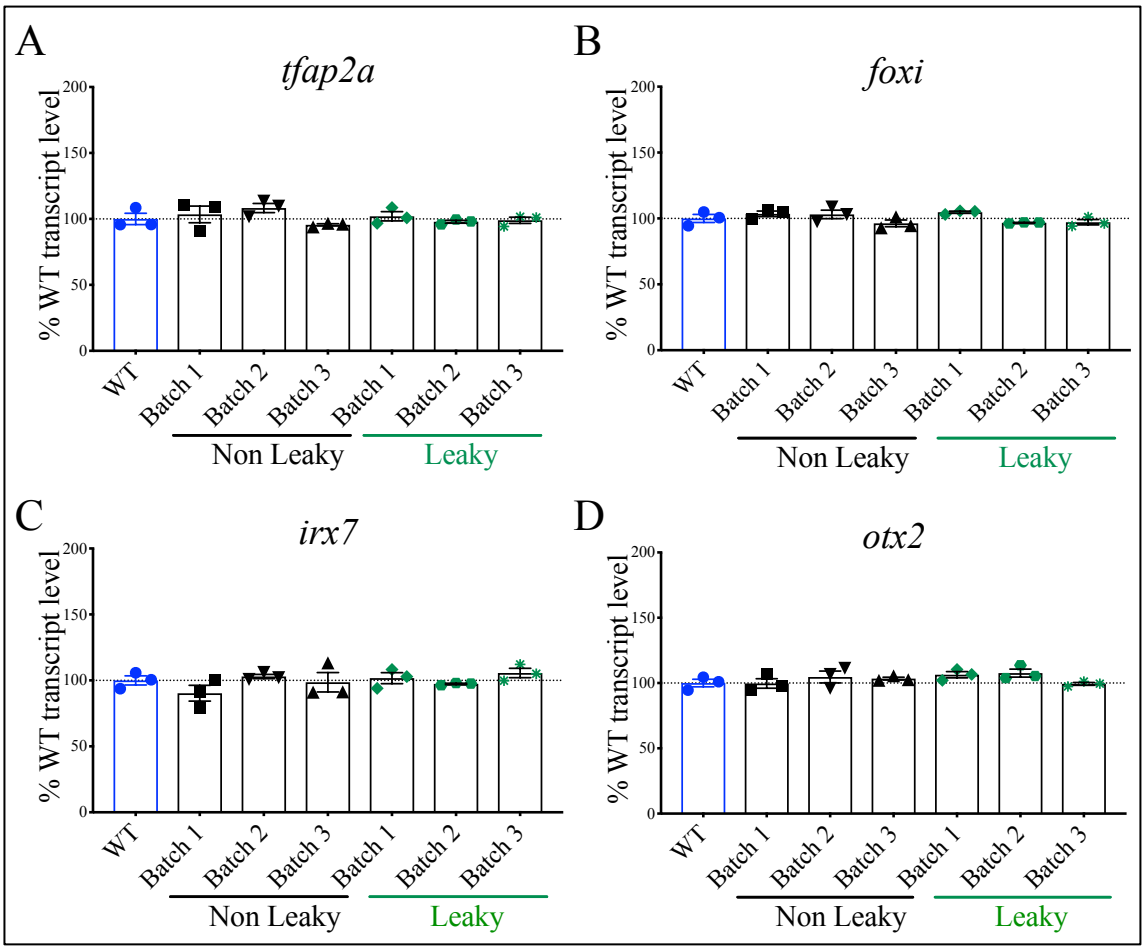


Figure 4.22 Expression of ectodermal genes in leaky vs non leaky embryos at 24.00 hpf. (A) – (D) No significant up or downregulation of any marker tested was observed. Reference gene *elf2* expression was used for normalisation. The assessment was based on three independently derived batches; each batch consists of three individual embryos; each point represents an embryo and is the mean of two technical replicates. Values are presented as % of WT expression ± SEM.

4.4.4 Correlation between *gfp* transcript level and aberrant gene expression in leaky embryos

As the misexpression of some endodermal and mesodermal genes in leaky embryos directly coincided with increased *gfp* expression, I next questioned whether these factors were directly correlated. I reanalysed the data for the genes that displayed statistically significant differences between leaky and non leaky embryos (*sox17, sox32, crcx4*, *mixl1* and *myf5*) and calculated a transcript ratio between target gene/*gfp* in leaky embryos. Strikingly, no difference was discernible between endodermal gene expression levels (*sox32, sox17*, *crcx4*); increased expression of these transcripts in leaky embryos was matched by an increased number of *gfp* transcripts with a ratio of almost 1:1. This suggested a positive correlation between *gfp* expression and the expression of these endodermal genes in the embryo (Figure 4.23). In

contrast, the mesendodermal and mesodermal genes, *mixl1* and *myf5* respectively, were downregulated in leaky embryos. Here, the increase in *gfp* transcripts negatively correlated with expression of these genes, with a ratio of 1:0.5 (2x downregulated) (Figure 4.23). Speculatively, this decrease could be explained by the concomitant increase in endodermal gene expression, if the lineage defining genes work to suppress one another during fate commitment. Conversely, the repression of these genes may be as an indirect result of the random integration of *sox17:gfp* construct into the genome *in loci* that would ultimately lead to repression of these genes when the *sox17* promoter is activated. Both these hypotheses however, remain to be tested.

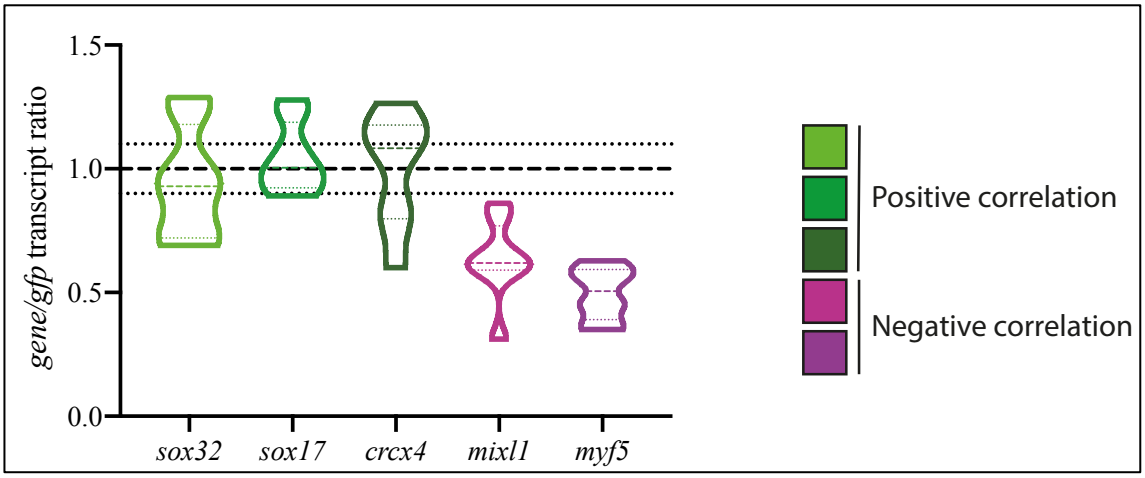


Figure 4.23 Violin plot of *gene/gfp* transcript ratio in leaky embryos. For each gene, expression was normalised to *gfp* expression. Genes in green show a positive correlation between number of gene transcripts and *gfp* transcripts. Genes in purple show a negative correlation, i.e. as the number of GFP transcripts increases, transcripts of these genes decrease. Dashed lines represent mean values, dotted lines denote 95% confidence intervals.

Taken together, these data suggested that the activity of the *sox17:gfp* promoter in this transgenic line, led to the misexpression of some early lineage markers expressed during gastrulation, and that the level of this misexpression was directly correlated to the level of *gfp* expression.

Interestingly, the misregulation of these genes did not result in any apparent phenotypical defects at 24 hpf, perhaps partially explained by the restoration of WT transcript levels by this stage as shown by RT-qPCR. However, at the protein level, the larger, ectopic domain of GFP was still apparent at 24 hpf (Figure 4.4). Individuals that had previously been shown to be leaky for GFP at 24 hpf, completely resembled the non leaky embryos by 48 hpf, with GFP expressed only in the pharyngeal arches (Figure 4.24C). This observation can probably be

explained by degradation of the ectopic GFP molecules between the time that expression of the mRNA ceases (sometime before 24 hpf) and the 48 hpf time point. The spatial domain of *sox17* transcripts was identical in non leaky and leaky embryos at 24 hpf and was detected only in the pharyngeal arches and hematopoietic precursors (Figure 4.24A, B), whereas GFP was still ectopically visible in leaky embryos (Figure 4.4). Thereafter, at 48 hpf, GFP was only observed in the spatial domain in which it should be expressed at that time point, i.e. the pharyngeal arches where *sox17* is truly expressed.

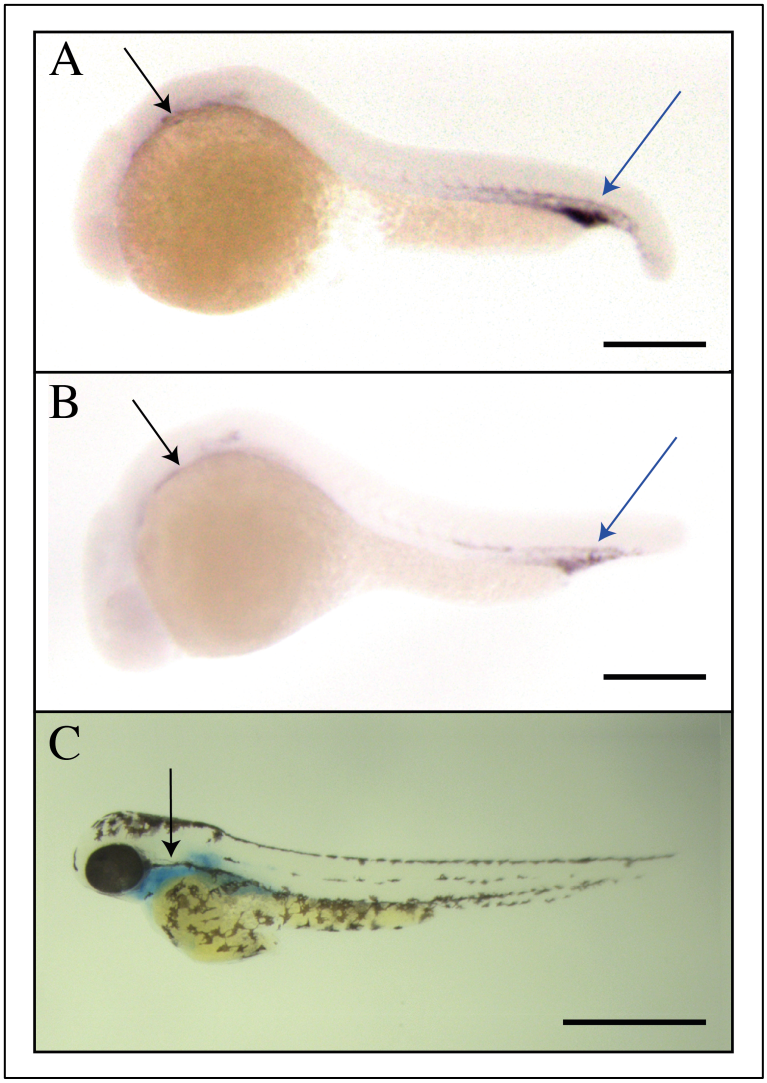


Figure 4.24 *gfp* transcript and GFP protein expression in leaky and non leaky embryos. *sox17* transcripts (WISH) were detected in the pharyngeal arches (black arrows), the dorsal aorta, the axial vein and hematopoietic precursors (blue arrows) in both non leaky (**A**) and leaky (**B**) embryos at 24 hpf. Note the overlapping *sox17* signal in A and B. (**C**) GFP (IHC) expression in leaky embryos resembled that of non leaky embryos at 48 hpf with domain of GFP restricted to the pharyngeal arches (black arrow). Lateral views anterior to the left. Scale bars represent 150 μm.

The restoration of all misexpressed genes in leaky embryos back to WT levels by 24 hpf presumably indicates a finetuning/compensatory mechanism. The phenomenon of genetic compensation in zebrafish following lesions at the level of the DNA has been previously described for a number of genes (Rossi et al., 2015). Similarly, the process of transcriptional adaptation, where compensatory mechanisms are employed following detection of mutant mRNA, has also been previously described (El-Brolosy and Stainier, 2017). However, most of these studies focus on compensatory mechanisms that occur following the loss of function of a gene; in contrast, my data in leaky embryos suggest that compensation is occurring primarily following the gain of function of genes, in particular *sox17*. This is investigated further in the Discussion chapter, where I attempt to identify possible reasons as to why leaky embryos, despite having severely misregulated endodermal gene expression, show no apparent phenotype.

4.5 Overview of gene expression time series

To gain a more comprehensive, overall view of the gene misregulation observed in leaky embryos, I decided to organise the entire dataset by developmental timepoint. In doing so, I was able to show the overall trends within the data. Previously, I observed trends in gene expression that, due to batching embryos into three samples of n=3 (as opposed to one batch of n=9), were not deemed to be statistically significant (see *tbxta* and *myf5* expression in Figure 4.15 for example). Embryos were batched according to the clutches they were collected from, but were spawned of the same parents. I therefore re-analysed the data combining all 9 embryos, with the only variable considered being whether the embryos were leaky, non leaky or WT (Figure 4.25). The data are presented % of WT transcript level ± SEM.

The time series data analysis for endodermal genes confirmed the results of the previous batch analysis with a 43.9% increase in *gfp* mRNA expression in leaky embryos compared to non leaky at 5.25 hpf ; by 7.00 hpf this had increased to 150% and at 9.00 hpf, *gfp* transcript expression in leaky vs non leaky embryos peaked at 346.2% (Figure 4.25).

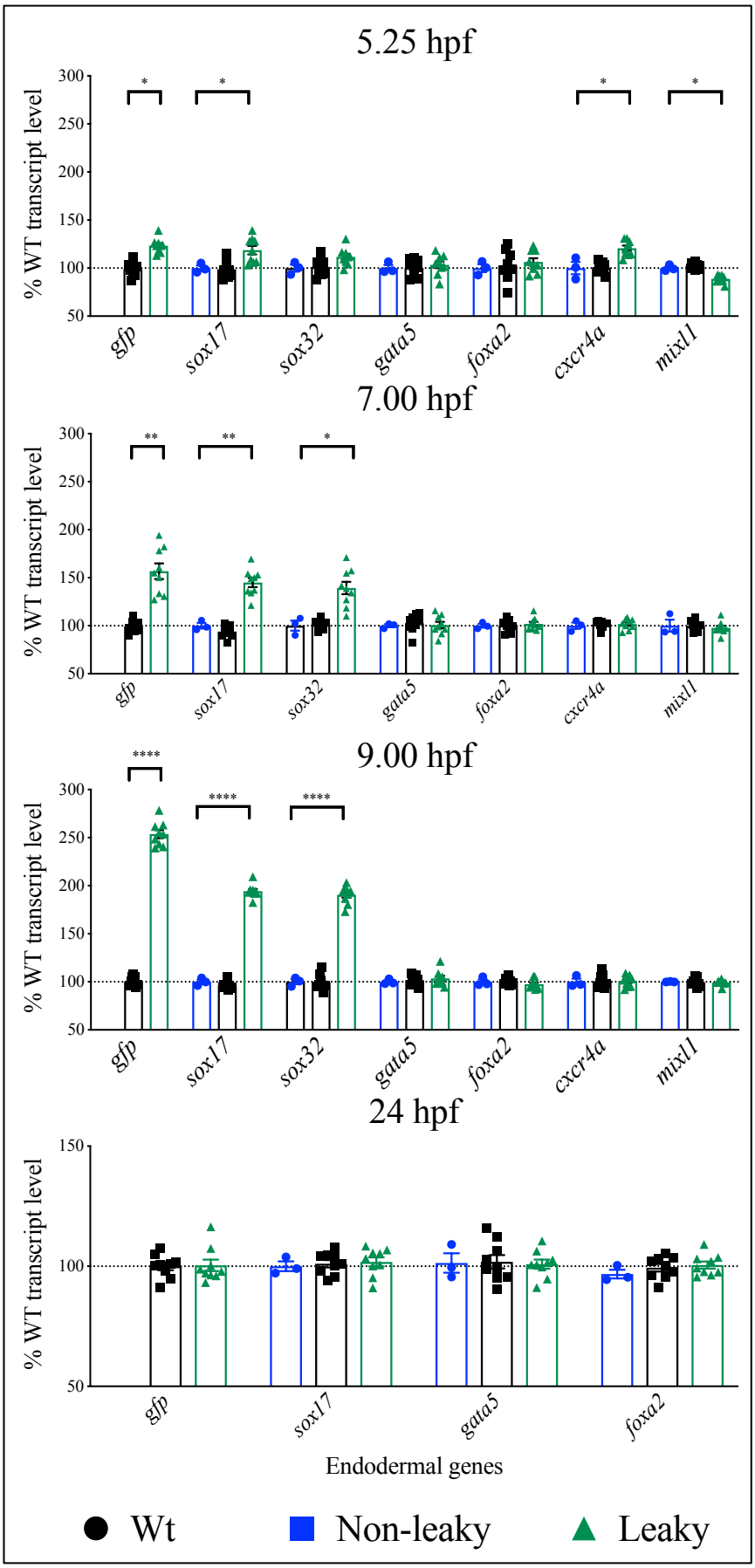


Figure 4.25 Temporal endodermal transcript quantification in developing WT, leaky and non leaky

embryos. Graph showing the mean value of nine biological replicates of normalised transcript abundance levels, expressed as a % of WT transcript level. Note the increase in *gfp*, *sox17* and *sox32* expression between 5.25 and 9.00 hpf. At the 5.25 hpf timepoint, *crcx4* is upregulated, and *mixl1 is* downregulated, however this observation did not persist beyond this stage. For all genes considered, no significant change was observed at 24 hpf. Bars represent SEM. Statistical analysis was performed using one-way ANOVA with Tukey's post-hoc test; *$p \leq 0.05$, **$p \leq 0.01$, ****$p \leq 0.0001$.

As I had previously identified changes in expression of mesodermal genes in leaky vs non leaky embryos, I decided to perform the same time series data analysis as described for endodermal genes above. These data showed that upregulation of *gfp* transcripts coincided with a decrease in *myf*5 expression by 8.7% at 5.25 hpf, 25.1% at 7.00 hpf and 22.2% at 9.00 hpf in leaky embryos, compared with the WT group (Figure 4.26). Pointedly, the trend towards reduced expression of *tbxta* that was visible but not significant when plotting each embryo batch separately (Figure 4.15) became significant with the new analyse (Figure 4.26, top row), confirming the validity of this approach (8.2 expression decrease %). Although there was a noticeable trend towards lower *tbxta* and *vox* transcript expression observed in leaky embryos compared to WT at 5.25 hpf, these changes did not reach statistical significance ($p = 0.06$) and did not persist following this time point. Notwithstanding, the transcript level of the other mesodermal genes tested was not affected in leaky embryos compared to WT.

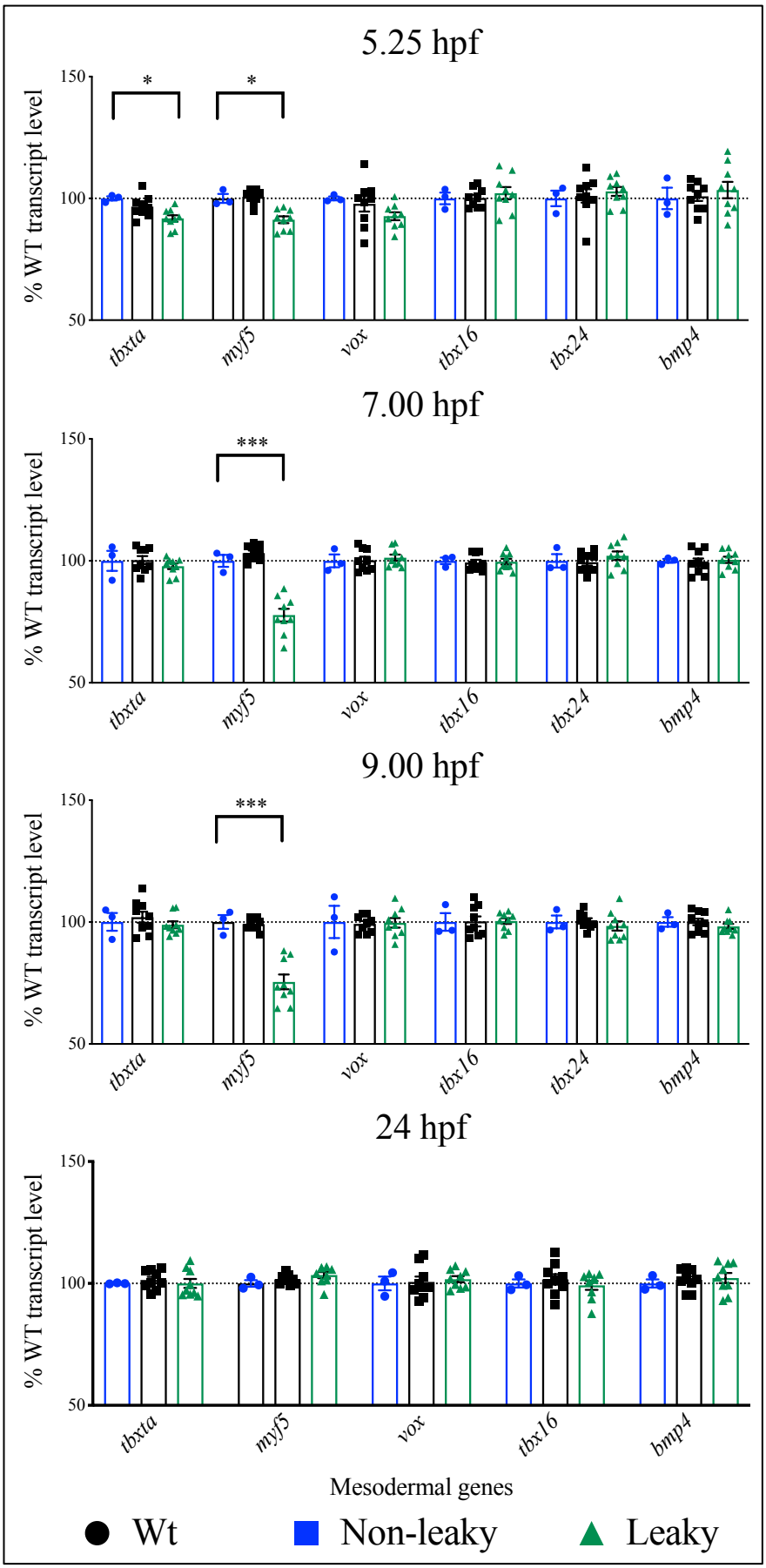


Figure 4.26 Temporal mesodermal transcript quantification in developing WT, leaky and non

leaky embryos. Bar graph showing the mean value of nine biological replicates of normalised transcript abundance levels, expressed as a % of WT transcript level. Note the marked decrease in expression of *myf5* and *tbxta* from 5.25 hpf onward. *vox* also showed a decrease in expression levels compared to WT, however this was not statistically significant. No changes were observed for any mesodermal genes in leaky embryos vs WT at 24 hpf. Bars represent SEM. Statistical analysis was performed using one-way ANOVA with Tukey's post-hoc test; *$p \leq 0.05$, ***$p \leq 0.001$

The same time series data analysis for ectodermal genes confirmed the results of the previous batch analysis: variation in the level of *gfp* transcripts did not coincide with any change in expression of any of the ectodermal genes tested, at any time point (Figure 4.27).

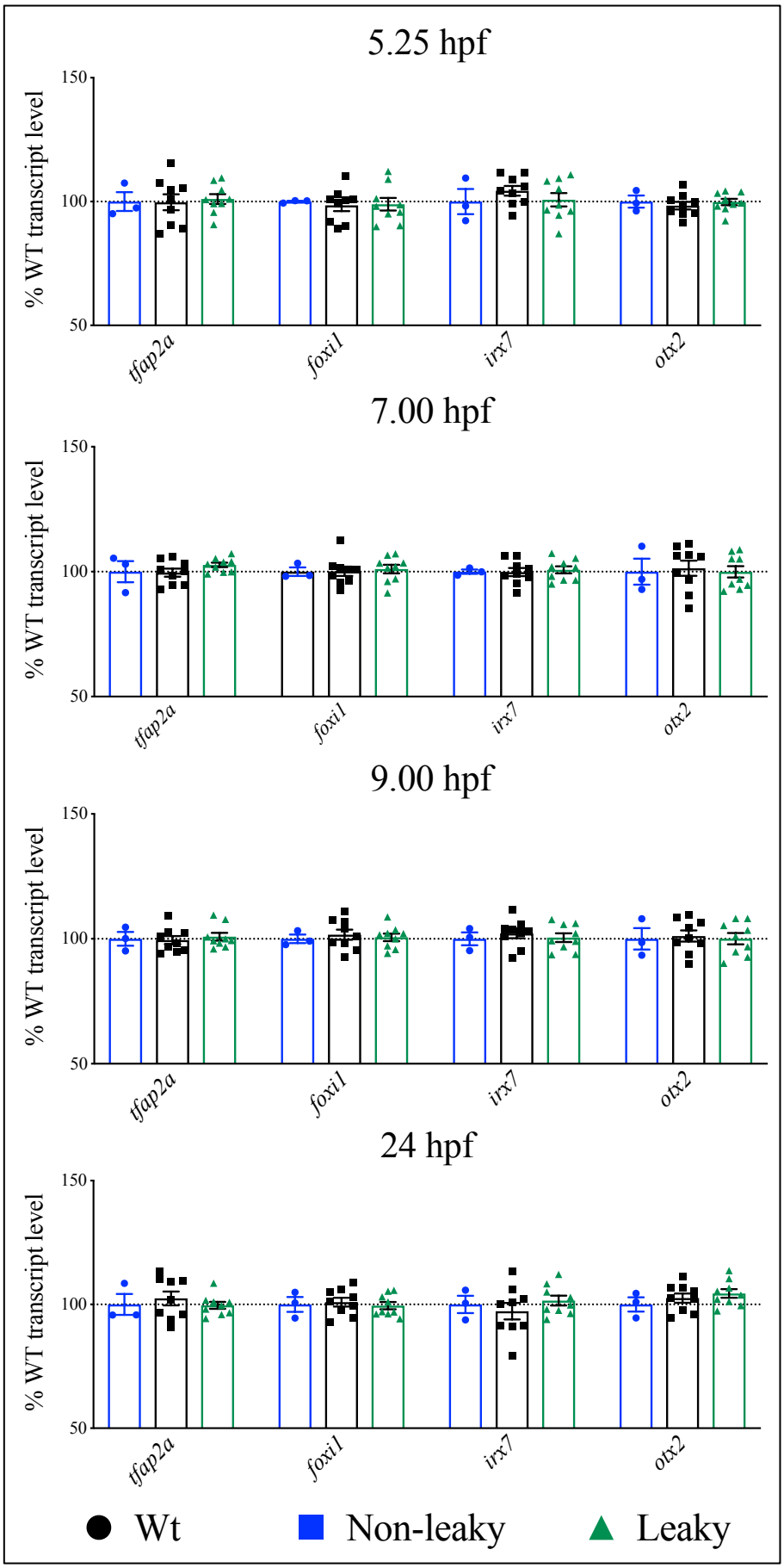


Figure 4.27 Temporal ectodermal transcript quantification in developing WT, leaky and non leaky

embryos. Bar graph showing the mean value of nine biological replicates of normalised transcript abundance levels, expressed as a % of WT transcript level. The upregulation of *gfp* expression in leaky embryos did not have any noticeable effect on ectodermal gene expression at all time point. Bars represent SEM.

Together, these data suggested that a misregulation of both endodermal and mesodermal genes was occurring in *sox17:gfp* leaky embryos. I therefore sought to further validate these findings, and compare the spatial expression domains of these genes, using WISH.

4.6 Spatial expression of misregulated genes

From my RT-qPCR data, I had identified genes that were upregulated or downregulated in *sox17:gfp* leaky embryos. As I used whole embryos for RT-qPCR, I could not be sure whether the expression changes I observed were due to increased mRNA expression in the same spatial domain as the wildtype embryos, or due to ectopic expression. To address this, I used WISH to examine the spatial expression domains of *sox17* and *myf5* in non leaky and leaky embryos. Leaky embryos have higher percentage of *sox17* transcripts and higher number of *sox17* positive cells (Figure 4.28). Furthermore, *myf5* expression in paraxial mesoderm segmental plate is downregulated at the end of gastrula stage (9.00 hpf) in leaky embryos (Figure 4.29).

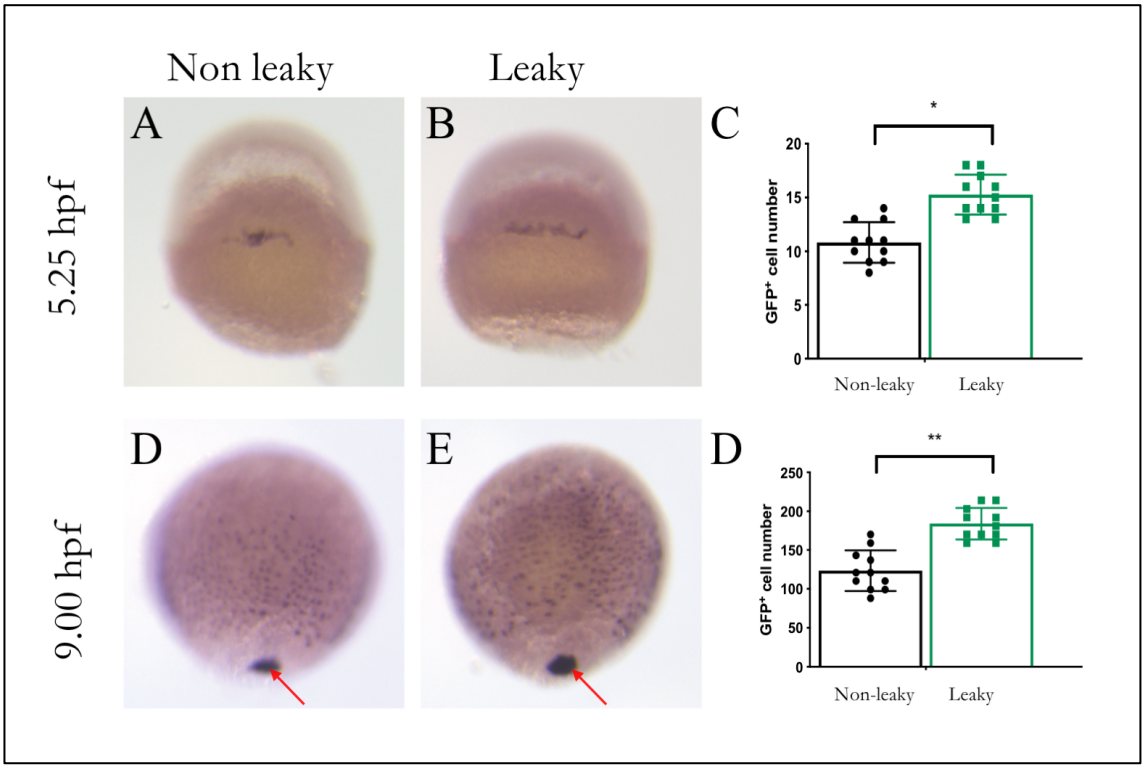


Figure 4.28 Leaky embryos display a higher number of sox17^{+} cells. *In situ* hybridization of non leaky embryos compared to leaky embryos for endogenous *sox17* expression (lateral views) at 5.25 hpf (**A**,**B**) and

9.25 hpf (**D**,**E**). Kupffer's vesicle indicated by arrows. (**C**) Quantification of cell numbers identified by WISH in panels A and B. (**D**) Quantification of cell numbers identified by WISH in panels D and E. *$p \leq 0.05$, **$p \leq 0.01$, Student's t test. Data are represented as mean ± SEM.

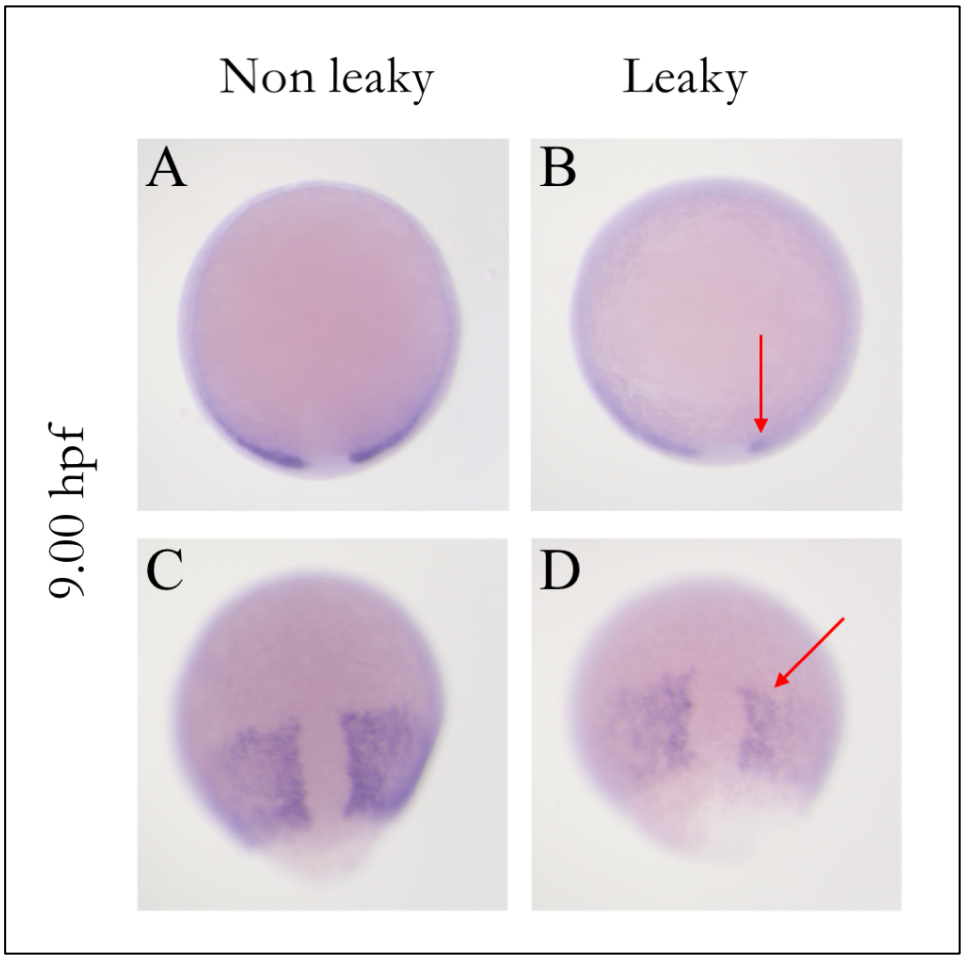


Figure 4.29 Leaky embryos show downregulation of *myf5* at 9.00 hpf. WISH of non leaky (**A**,**C**) embryos compared to leaky embryos (**B**,**D**) for *myf5* expression showed decreased expression in paraxial mesoderm in leaky embryos (red arrows). A and B are posterior views, dorsal down; C and D are lateral views.

4.7 *sox17:GFP* flow cytometry to isolate endodermal cells

As discussed above, as I used whole embryos for RT-qPCR, I could not be sure whether the altered expression of endodermal and mesodermal genes in leaky embryos occurred only in endodermal cells or, if the leaky expression was causing gene misregulation in other tissues. The WISH performed above went some way to address this question, showing downregulation of *myf5* in mesodermal cells and upregulation of endodermal cells characterised by *sox17* expression in leaky embryos. However, I sought to further investigate this by dissociating the embryos and isolating GFP^+ cells using fluorescence-activated cell sorting (FACS). Ultimately, I intended to define the endodermal transcriptomic signature using endodermal cells isolated

from this line, using non leaky embryos. Once the FACS protocol was optimised (see below and Materials and Methods), it allowed me to directly assess gene expression in the GFP labelled cells from non leaky embryos, and then proceed to assess differentially expressed genes in GFP subpopulations isolated from both non leaky and leaky embryos. A summary of the optimised process I used to isolate GFP^+ cells from embryos of the transgenic line is depicted in Figure 4.30. This methodology was applied first to non leaky embryos and then to leaky embryos to enable me to isolate the subpopulations for further analysis. I selected embryos at the 9.00 hpf time point as this was when the GFP fluorescence was the strongest (during gastrulation) and the previously observed differences in endodermal gene expression were the most significant at this stage.

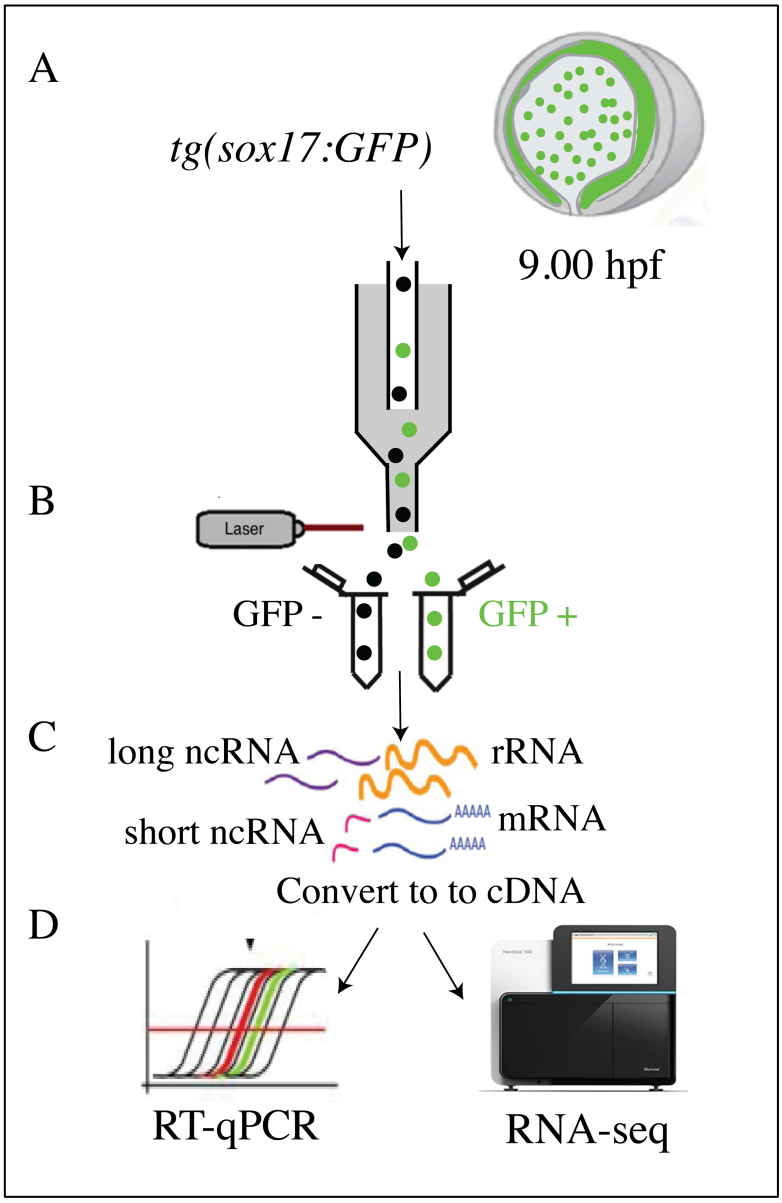


Figure 4.30 Flow cytometry workflow to isolate GFP+ cells from the *sox17:GFP* line. (A)

Mechanical dissociation of *sox17:GFP* embryos at 9.00 hpf. (**B)** Cells were sorted into $GFP^+$and GFP^- populations, using a FACS Aria II Flow Cytometer (**C**) RNA was isolated using TRIzol as previously described and (**D**) used for RT-qPCR or RNA-seq library preparation.

4.7.1 Data analysis using flow cytometry

Flow cytometry is a widely used technique that employs laser-based technology to count, sort, and profile heterogenous mixtures of cells according to their physical and chemical characteristics. During flow cytometry, cells are suspended in a liquid stream and are then passed, one at a time, through a laser light beam, and their interaction with the light is measured. Fluorescence activated cell sorting (FACS) is a specialised form of flow cytometry that facilitates physical separation of cells of interest from a heterogenous cell sample into two or more populations according to specific light scattering, emitted fluorescence and other user specified parameters. Every cell that passes through the laser is detected, counted as a distinct event and channelled to a specified tube for collection. In this way, FACS has been used to sort fluorescent cells from a heterogenous population that originated from a transgenic reporter line, thus isolating cells of a particular tissue or lineage (Gallardo and Behra, 2013; Manoli and Driever, 2012; Rougeot et al., 2014; Singh et al., 2018).

Acquisition of my flow cytometry data and the subsequent isolation of different GFP subpopulations were based on an optimised gating strategy. Gates are regions of cells populations with similar properties that are defined by the user. The first step of this process was to group populations of cells based on their light scattering properties. These were forward-scatter light (FSC), side-scatter light (SSC) and dye-specific fluorescence emission signals. These former measurements (FSC and SSC) are used to describe where the light is collected in respect to the path of the laser; FSC measures light refracted by cells that travels in the same direction as the original light path, SSC measures light refracted by cells that travels in a different direction than the original light path (measured at a 90° angle to the excitation line). Forward and side scatter measurements estimate the size and granularity of the cells respectively. These light scatter patterns are able to distinguish cellular debris and dead cells within the sample, as these have the lowest level of forward scatter and are found at the bottom left corner of the density plot (Figure 4.31). This is because live cells are larger and thus show high SSC and FSC values; debris and dead cells have low SSC and FSC values and can therefore be separated into different populations based on these values alone. In my

experiments, I increased the forward scatter threshold to prevent the collection of these events and consistently obtained more that 50% of alive cells.

In addition to setting the appropriate gates, controls were essential to reliably discriminate the populations of interest from background noise. I adopted two critical controls to ensure reproducibility of the sorting data: I used WT embryos and DAPI staining to check i) any morphological differences or alterations in cell size/granularity between the WT and *sox17:GFP* line and ii) how to compensate for autofluorescence.

Intensity of the GFP dye-specific fluorescence signal was used to sort pure GFP^+ cell populations, but I identified that the level of autofluorescence within the whole sample was an issue. The presence of dead cells is known to affect the sorting process and therefore the quality and consistency of the data (Cossarizza et al., 2017). This is because dead cells have greater autofluorescence than live cells, leading to false positives. I have already described the gates I set based on the FSC and SSC measurements to help remove debris and dead cells, and whilst this alleviated most of the problem of dead cells, it did not completely eliminate it. Cells have a natural level of fluorescence (autofluorescence) which can be due to the presence of collagen and elastin, cyclic ring compounds such as NADPH, FAD and riboflavin, aromatic amino acids and/or cellular organelles such as mitochondria and lysosomes (Andersson et al., 1998). DAPI emits blue light at a wavelength of 461 nm and therefore stands out in contrast to the fluorophore of interest, GFP, that emits green light at 510 nm. Dead cells have a permeabilized cell membrane that increases their accessibility to DAPI, as a result, dead cells have much higher fluorescence than living cells, where the cell membrane is intact. I therefore used DAPI staining to set a second gate to determine the level of background fluorescence/ autofluorescence (Figure 4.31).

Dead cells take up DAPI readily, but it can cross the intact membrane of live cells at a lower rate, causing false positive staining if left on the cells for too long. I found the optimum staining time to only label dead cells to be between 3 to 10 minutes; no discernible differences in staining were observed during this timeframe, however longer incubation times increased the DAPI positive population significantly.

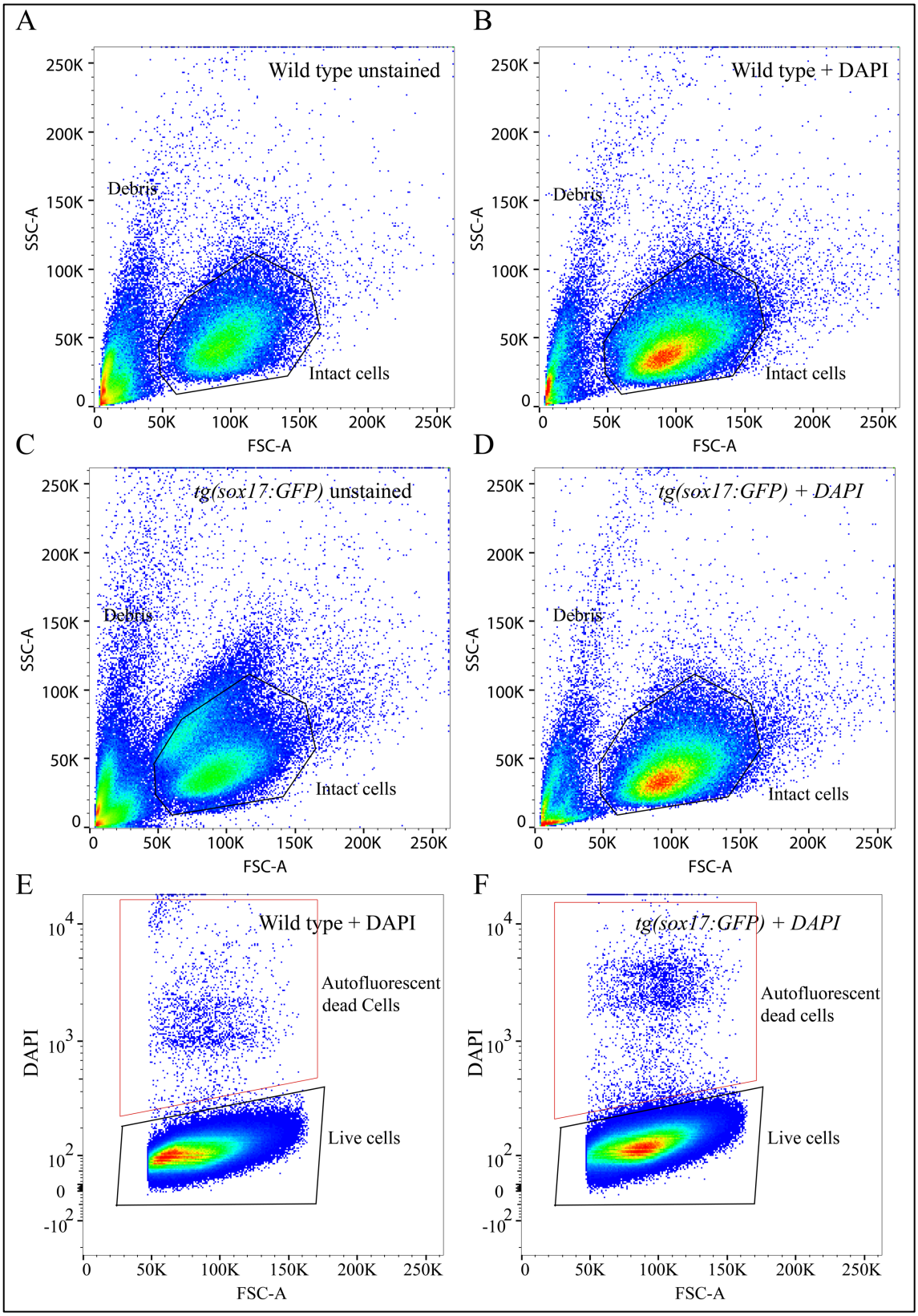


Figure 4.31 Unstained WT and *sox17:GFP* viability controls. (**A-D**) Use of FSC and SSC gating to remove debris/dead cells. SSC vs FCS density plot for (A) WT, (B) WT and DAPI, (C) *sox17:GFP* and (D) *sox17:GFP* and DAPI. (**E-F**) Use of DAPI and FSC gating to specifically exclude dead cells (red polygons) from the populations sorted in (C) and (D) respectively. Each dot on the plot represents an individual event.

The processes involved in identifying GFP positive cells are described below; briefly, once the cell population was selected by its forward and side scatter characteristics and negative gates were appropriately set, I sought to identify where the WT population was in respect to the *sox17:GFP* cells in terms of fluorescence. Therefore, to determine the level of autofluorescence, I plotted the DAPI channel against the GFP channel. The plot was then divided into regions depending on the intensity of the fluorescence (*y*-axis was DAPI intensity and *x*-axis was GFP intensity), to create a series of extraction gates. In particular, a vertical line drawn just above GFP intensity at 2 x 10^2 identified all WT cells (Figure 4.31 A) and allowed me to separate GFP^+ cells (Figure 4.32A). I also plotted the data in a single dimension in a univariate histogram, where the *y*-axis was the number of events (cell count) and the *x*-axis was relative fluorescence intensity detected in the GFP channel. I observed a single distinct peak (a large number of events detected at one particular intensity) in the WT histogram, at between 0 and 2 x 10^2 and I interpreted this as the negative dataset (representing mesodermal and ectodermal cells). Consistent with this, the *sox17:GFP* histogram showed a population of cells above 2 x 10^2, the GFP^+ endodermal cells. The GFP^+ population became more easily visible when the WT control histogram (red) was overlaid onto the *sox17:GFP* histogram (blue), allowing the shoulder of positive GFP cells to be accurately identified (Figure 4.32F).

The proportion of cells in each population was then quantified accordingly (Figure 4.32). In this case, the bottom left quadrant represented WT cells and the bottom right quadrant represented GFP^+ cells. In the WT cell sample, 100% of cells fell into the WT quadrant; for the *sox17:GFP* line, 77.2% fell into the WT quadrant, with the remaining 22.8% were GFP^+.

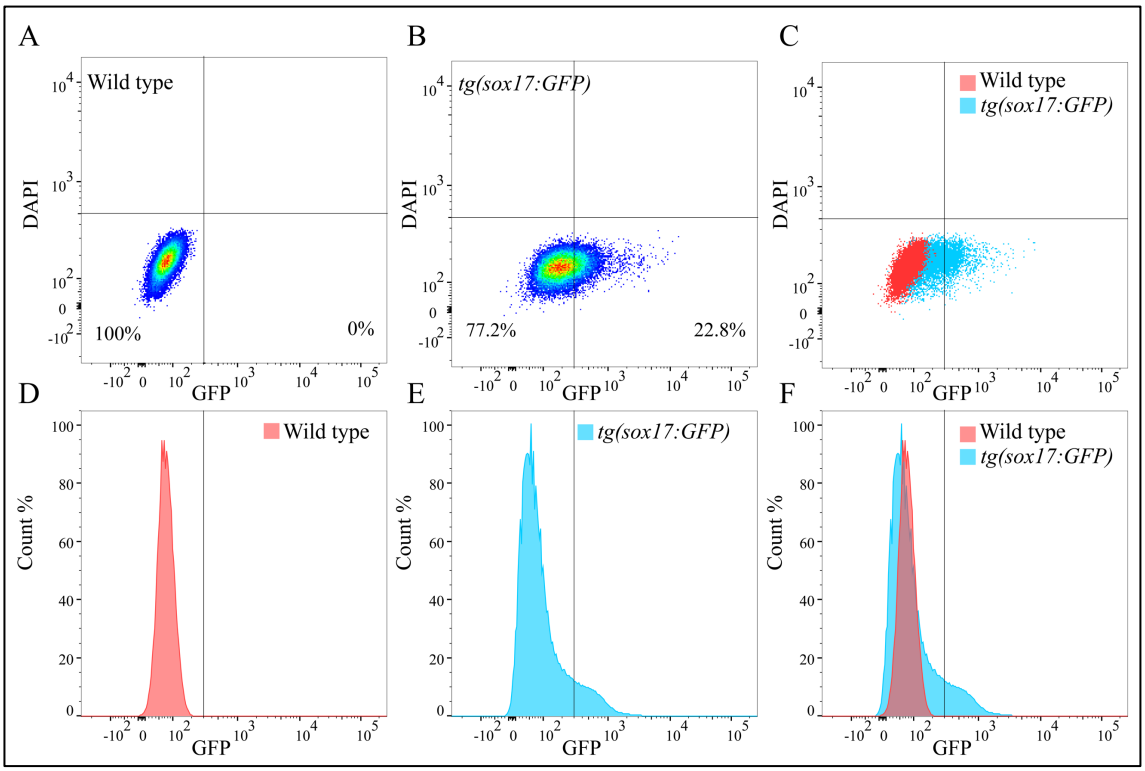


Figure 4.32 Determination of the negative gating strategy. Cells from WT embryos were used to define the negative population. (**A**) Single parameter histogram defining the negative (WT) population. (**B**) as (A) except with the *sox17:GFP* sample. 22.8% of cells were GFP+. (**C**) Merge of A and B. (**D**) and (**E**) One dimensional univariate histograms showing the data in (A) and (B) respectively, plotted as cell count vs GFP intensity. Note the shoulder of GFP+ cells in (E). (**F**) Merge of D and E.

From these data, I concluded that there were no significant modifications required in the forward and side scatter profiles I had set between WT and the *sox17:GFP* line. I therefore proceeded considering the WT as the known negative control, and set the negative gates around the WT cell population accordingly, allowing the isolation of the GFP^+ cells. Additionally, the DAPI staining improved the accuracy of the gating strategy by removing extra dead cells that were not excluded earlier based on the FSC/SSC ratio.

4.7.2 Gating strategies

To minimize the collection of unwanted GFP^+ non endodermal cells, I now adopted a sequential gating strategy to sort *sox17:gfp* cells. Here, each plot had two measurement parameters, one on the *x*-axis and one on the *y*-axis, and the events (cell counts) displayed as a density (or dot) plot. Gating is a data reduction technique; the principle is to continue applying more stringent parameters to remove debris and other events that are not of interest. As gating proceeded, fewer events within each gate were retained, thus demonstrating the

importance of collecting sufficient numbers of cells to run both diagnostic RT-qPCR and RNA-seq library preparation. I observed that with 50,000 cells I was able to obtain both sufficient RNA yields and stable average gene expression; however, this was not possible with lower numbers of cells (30,000 to 50,000). Briefly, live cells were determined by forward and side scatter as previously described (Figure 4.33A) and then the population was selected for by height over size (the height or width against the area for forward scatter or side scatter) thus identifying singlets from doublets (Figure 4.33B). Doublets need to be excluded as they could potentially consist of one GFP^+ cell and one GFP^- cell. Doublets have double the area and width values of single cells whilst the height remains approximately the same. Therefore, disproportions between height, width and area were used to identify doublets. Live cells were again identified and gated by their DAPI expression (Figure 4.33C) and then the GFP^+ cells were identified and gated using the expression parameters previously set to identify the WT population (Figure 4.33D and E) to distinguish between high GFP/GFP^+ and low GFP/ GFP^- expressing cells.

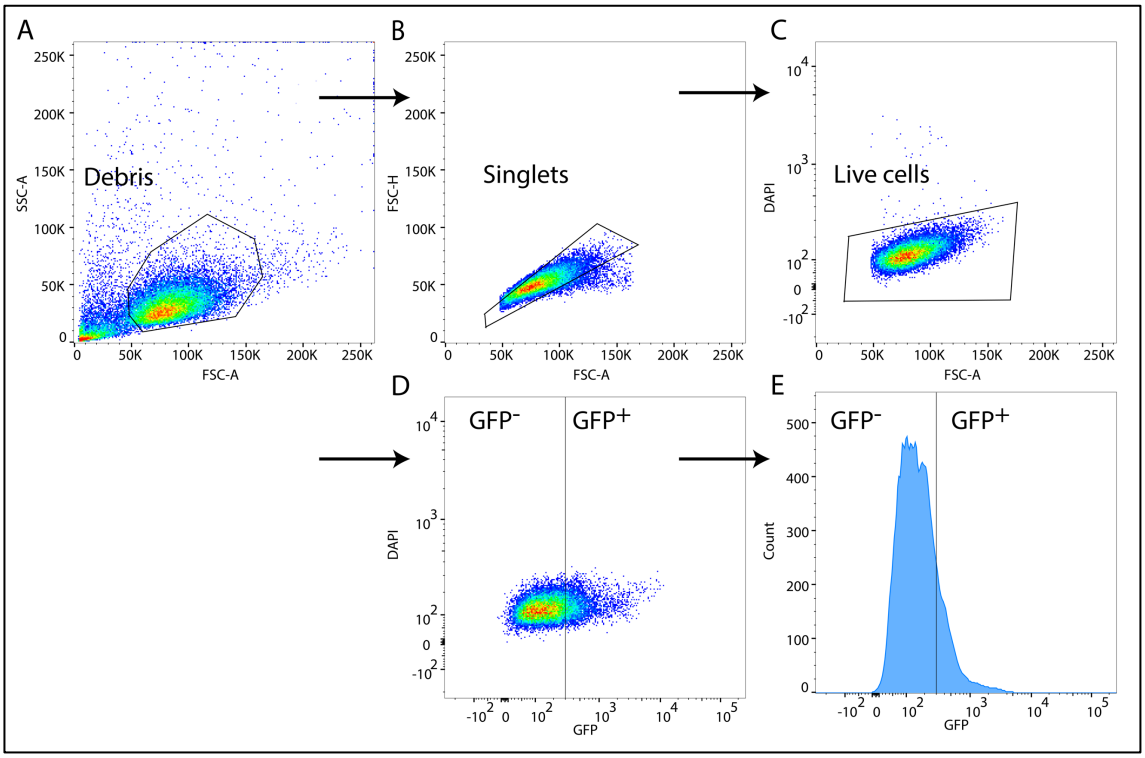


Figure 4.33 Sequential gating strategy to isolate GFP^+ expressing cells. Fluorescence activated cell sorting of GFP^+ cells previously sorted from 9.00 hpf *sox17:GFP* embryos. Gates are shown as black polygons. (**A**) Cells were subject to forward scatter (FSC-A) and side scatter (SSC-A) analysis to remove dead cells/debris as previously described. (**B**) Sorted cells from (A) were then sorted to isolate singlets only, based on area:height ratio and (**C**) viability as determined by DAPI. (**D**) and (**E**) Intact, viable cells were then sorted

based on fluorescence intensity into a GFP^- population (intensity < 2 x 10^2) and a GFP^+ population (intensity > 2 x 10^2).

This strategy allowed me to isolate a relatively pure population of GFP^+ cells from non leaky embryos from the *sox17:GFP* line, to be used in downstream analysis of the endoderm transcriptomic signature.

4.7.3 FACS of non leaky and leaky embryos

In order to try and explain the misregulation of genes identified in leaky embryos vs non leaky embryos in whole embryo RT-qPCR, I then proceeded to apply the aforementioned gating strategy to leaky embryos. Firstly, I used a pseudo-colour density plot to show the overall trend of the cells in non leaky vs leaky embryos (Figure 4.34A-C). Quadrants were defined to delineate GFP^+ from WT cells as previously described, and the proportion of cells in each quadrant was recorded. Leaky embryos showed 2.5x more cells in the GFP quadrant than non leaky embryos (63% compared to 25%).

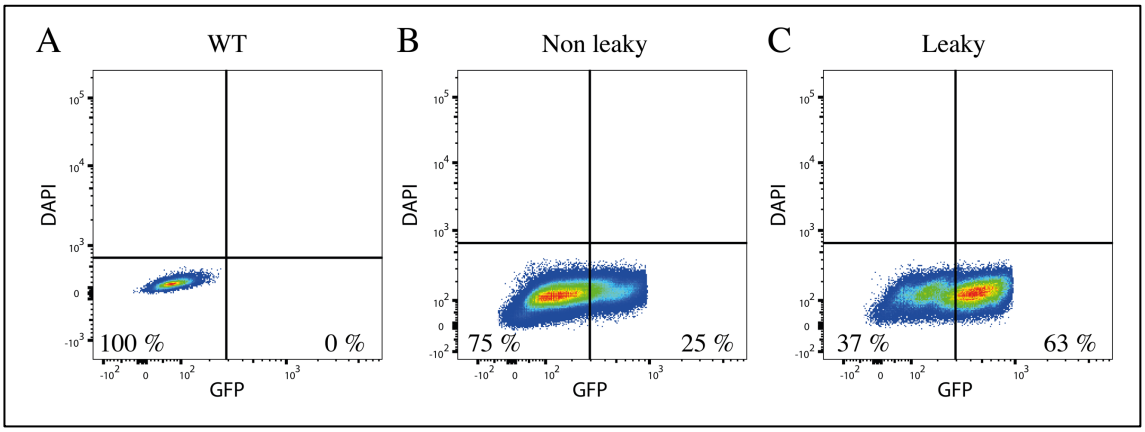


Figure 4.34 Pseudo-colour density plots showing the percentage of GFP^+ cells in non leaky and leaky embryos. DAPI intensity is plotted against GFP intensity for cells from (**A**) WT, (**B**) non leaky and (**C**) leaky *sox17:GFP* embryos, all at 9.00 hpf. Note that significantly more cells from leaky embryos fEll into the GFP^+ quadrant (bottom right) compared to non leaky.

To better depict the differences between leaky and non leaky embryos in terms of GFP fluorescence, I then proceeded to compare the data using a univariate histogram plot. The overlay clearly showed how the bulk number of cells from non leaky embryos showed low GFP intensity (< 2 x 10^2) whereas the majority of cells from the leaky embryos showed high GFP intensity values (> 2 x 10^2), shifting the plot to the right (Figure 4.35).

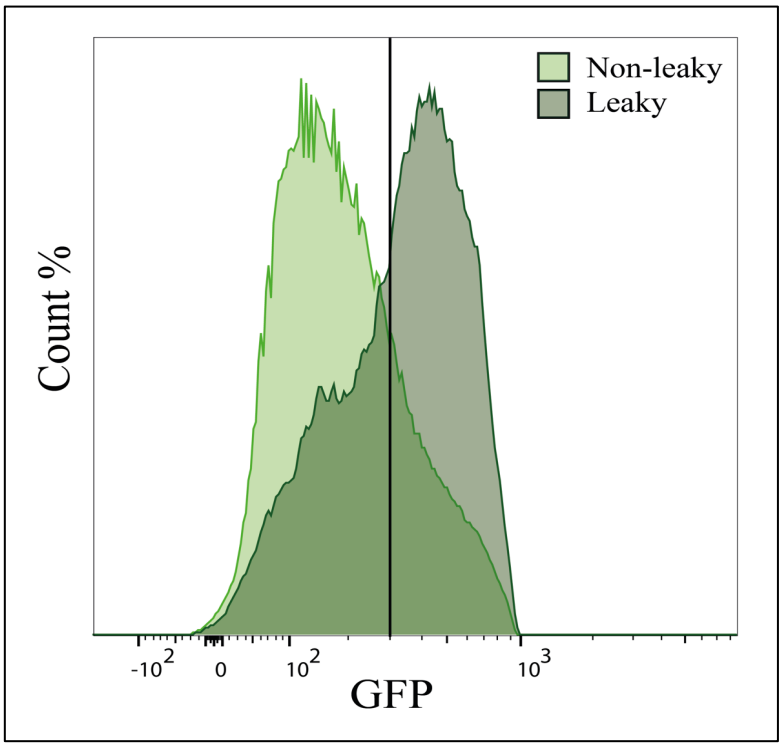


Figure 4.35 Overlay of univariate histograms for non leaky (light green) and leaky (dark green) embryos. Cell count was plotted against GFP intensity; note the shift to the right in cells from the leaky embryos.

I then proceeded to collect four biological replicates of 50,000 GFP^- expressing cells and 50,000 GFP^+ expressing cells each for both conditions (non leaky and leaky), in order to compare relative expression levels of some of the previous germ layer markers. As shown in Figure 4.36, each biological replicate was consistent within its own condition (either non leaky or leaky) with the exception of replicate 3 which showed an anomalous profile in both conditions. Replicate 3 in the non leaky profile showed a bigger bump of cells at a GFP intensity of $> 2 \times 10^2$ compared to the other replicates; replicate 3 in the leaky profile showed an overall higher GFP intensity than the other replicates.

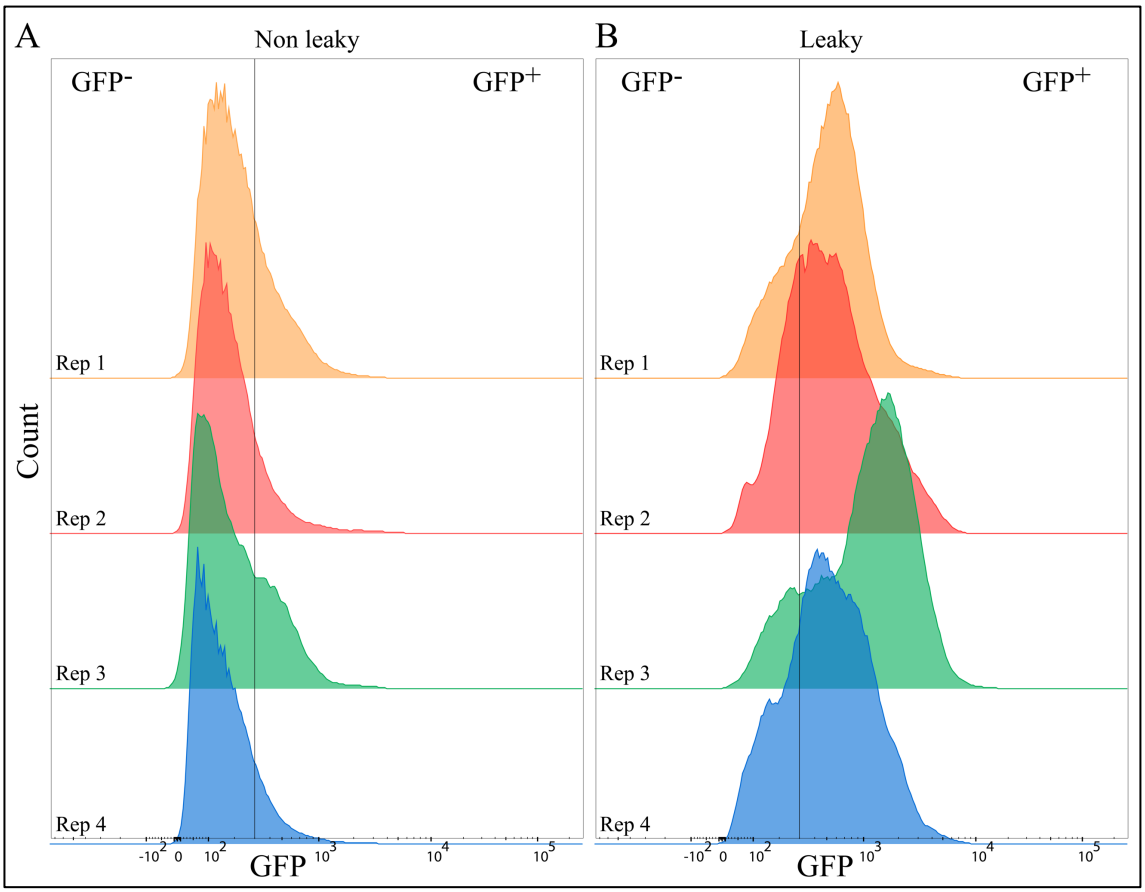


Figure 4.36 Flow cytometry analysis of four biological replicates for non leaky and leaky embryos. GFP intensity profile plotted against cell count of GFP^+ cells isolated from either (**A**) non leaky, or (**B**) leaky *sox17:GFP* embryos at 9.00 hpf. Plots are color-coded for biological replicates; embryos from the same clutch were segregated into non leaky and leaky conditions, before dissociation and FAC-sorting. Black vertical lines separate GFP^-/GFP^+ expressing cells.

I next quantified the relative proportions of low GFP and high GFP expressing cells and reported the percentage in the respective condition, either non leaky or leaky. The average proportion of cells expressing low GFP across the four replicates was 83.5% in non leaky and 22.7% in leaky. Concomitantly, the average across the four replicates for high GFP expressing cells was 16.5% in non leaky, compared to 77.3% in leaky (Figure 4.37).

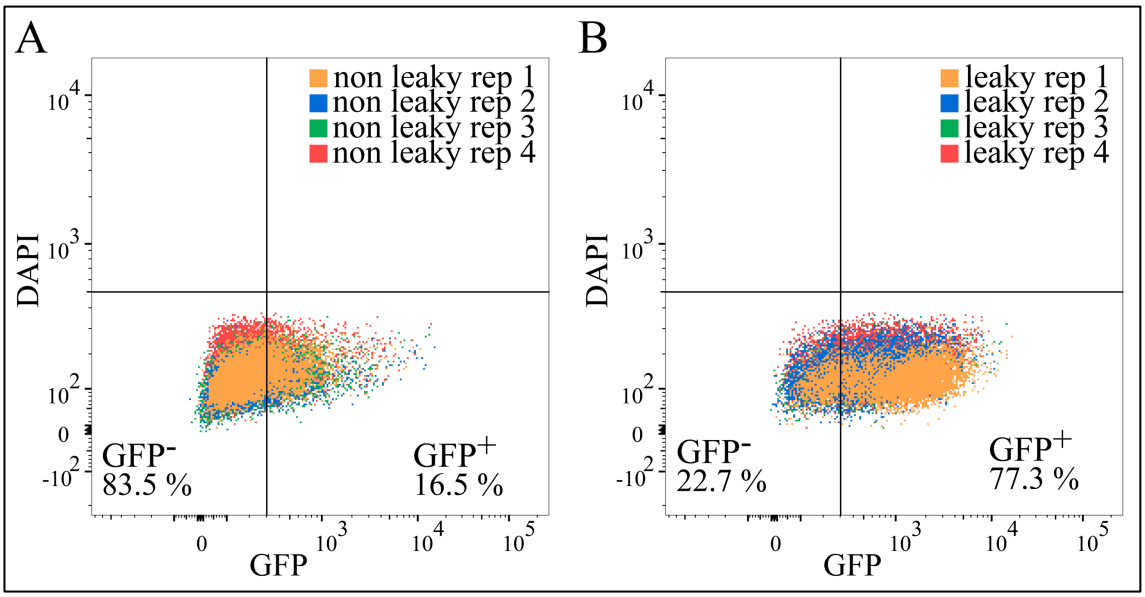


Figure 4.37 DAPI vs GFP intensity plots of the biological replicates for non leaky and leaky embryos. Overlap of the 4 biological replicates for (**A**) non leaky embryos and (**B**) leaky embryos. Note the reciprocal pattern of high % of GFP^- expressing cells/low % of GFP^+ expressing cells in the non leaky condition and vice versa for the leaky condition.

In addition to counting the number of cells and then determining the proportions of cells within the gates, I also calculated measurements and statistics for many other parameters to aid my analysis (Table 4.2). In particular, I determined the mean fluorescence intensity (MFI) for non leaky and leaky embryos which clearly showed the differences in fluorescence for the two conditions, with the MFI of leaky embryos being almost double that of non leaky embryos Table 4.2 and Figure 4.38).

Table 4.2. Summary of flow cytometry statistics for non leaky and leaky embryos

Non leaky embryos	Cell subsets (SSC-A/FCS-A plot)	Singlets subset (FSC-H/FSC-A plot)	DAPI subset (DAPI/FSC-A plot)	GFP^- cells	GFP^+ cells	Mean (GFP^-)	Mean (GFP^+)
Rep 1	74.9	91.5	99.7	81.4	18.6	141.0	674.0
Rep 2	55.2	84.0	93.0	83.6	16.4	145.0	709.0
Rep 3	54.8	94.4	99.1	71.0	29.0	151.0	607.0
Rep4	56.4	94.4	99.4	84.3	15.7	136.0	663.0
Mean (n=4)	60.3	91.1	97.8	80.1	19.9	143.3	663.3
SEM	4.9	2.5	1.6	3.1	3.1	3.2	21.2

Leaky embryos	Cell subsets (SSC-A/FCS-A plot)	Singlets subset (FSC-H/FSC-A plot)	DAPI subset (DAPI/FSC-A plot)	GFP^- cells	GFP^+ cells	Mean (GFP^-)	Mean (GFP^+)
Rep 1	67.4	91.6	99.6	22.1	77.9	150.0	1083.0
Rep 2	61.7	83.8	93.2	24.7	75.3	202.0	1001.0

Rep 3	46.9	95.4	99.1	13.2	86.8	188.0	1727.0
Rep4	48.0	94.1	98.5	26.3	73.7	177.0	924.0
Mean (n=4)	56.0	91.2	97.6	21.6	78.4	179.3	1183.8
SEM	5.1	2.6	1.5	2.9	2.9	11.0	184.0

Predictably, the low GFP expressing population of cells accounted for a higher percentage of cells in non leaky embryos compared to leaky and vice versa. (Figure 4.38).

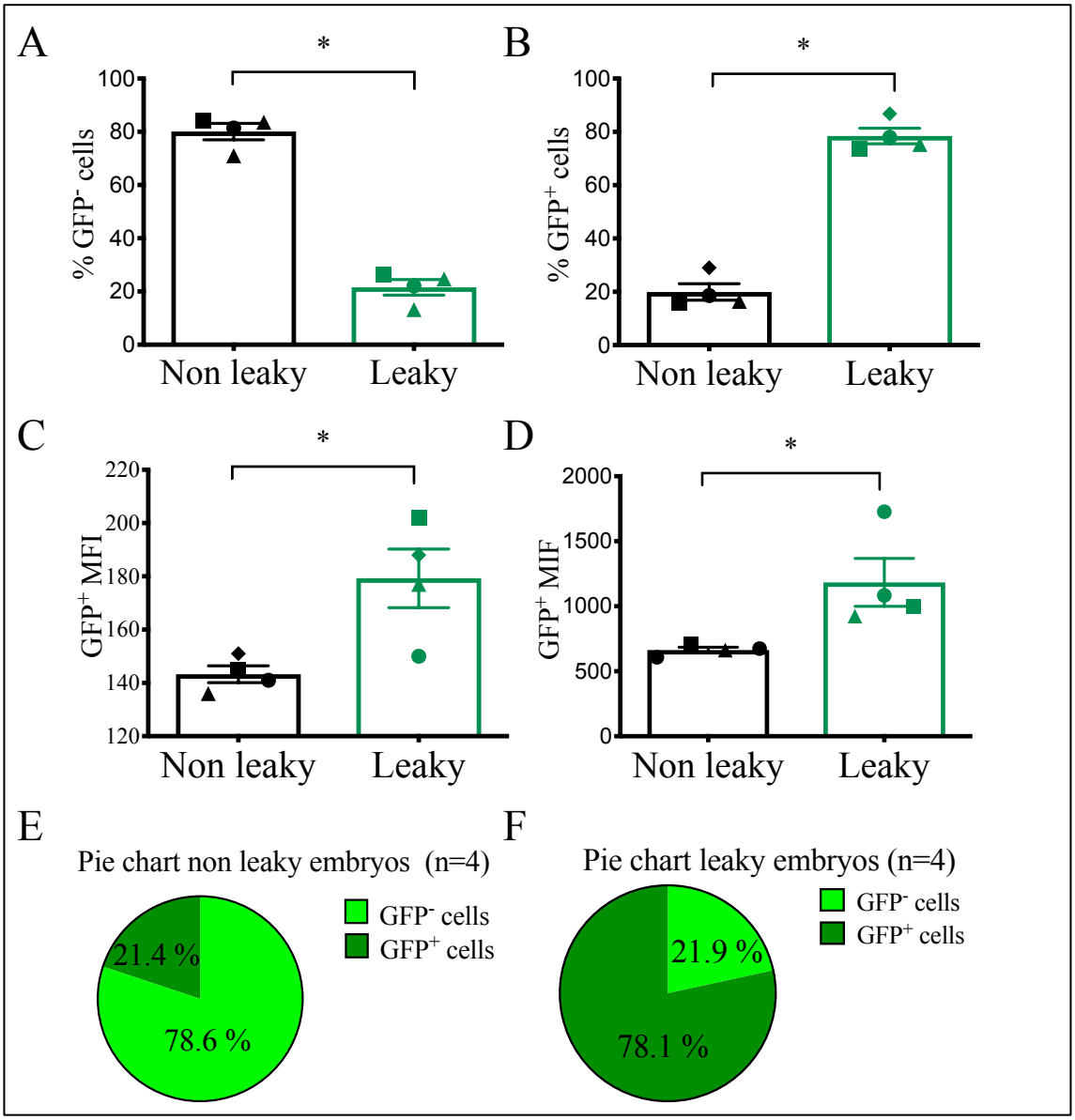


Figure 4.38 Flow cytometry measurements for leaky and non leaky embryos. % of GFP^- and GFP^+ expressing cells for (**A**) non leaky embryos and (**B**) leaky embryos. MFI for non leaky and leaky embryos in GFP^- (**C**) and GFP^+ (**D**) cell populations. Summary pie charts of subpopulation in (**E**) non leaky embryos and (**F**) leaky embryos. Statistical analysis was performed using Wilcoxon T test: $*p \leq 0.05$ (n = 4, mean ± SEM).

4.7.4 Gene expression analysis (RT-qPCR) in non leaky vs leaky GFP$^+$ cells

Having isolated GFP$^+$ cells from both non leaky and leaky *sox17:GFP* embryos at 9.00 hpf, I next performed RT-qPCR for specific known endodermal, mesodermal and ectodermal markers. In order to do so, cDNA was synthesised from total RNA extracted from GFP$^+$ expressing cells and GFP$^-$ expressing cells from both non leaky and leaky embryos. Transcript levels in GFP$^+$ expressing cells, from both non leaky and leaky embryos, were then expressed as a fold change of transcript levels in GFP$^-$ expressing cells (Figures 4.39, 4.40 and 4.42).

FAC sorted GFP$^+$ cells from non leaky embryos showed high expression of endodermal markers (Figure 4.39 yellow bars) whereas levels of markers of mesoderm (Figure 4.39 red bars) and ectoderm (Figure 4.38 blue bars) were enriched in GFP$^-$ sorted cells. As expected, *sox32* and *sox17* transcripts were readily detected in the GFP$^+$ cells, suggesting that endodermal markers were enriched in the GFP$^+$ population I selected, which were therefore presumably endodermal progenitor cells.

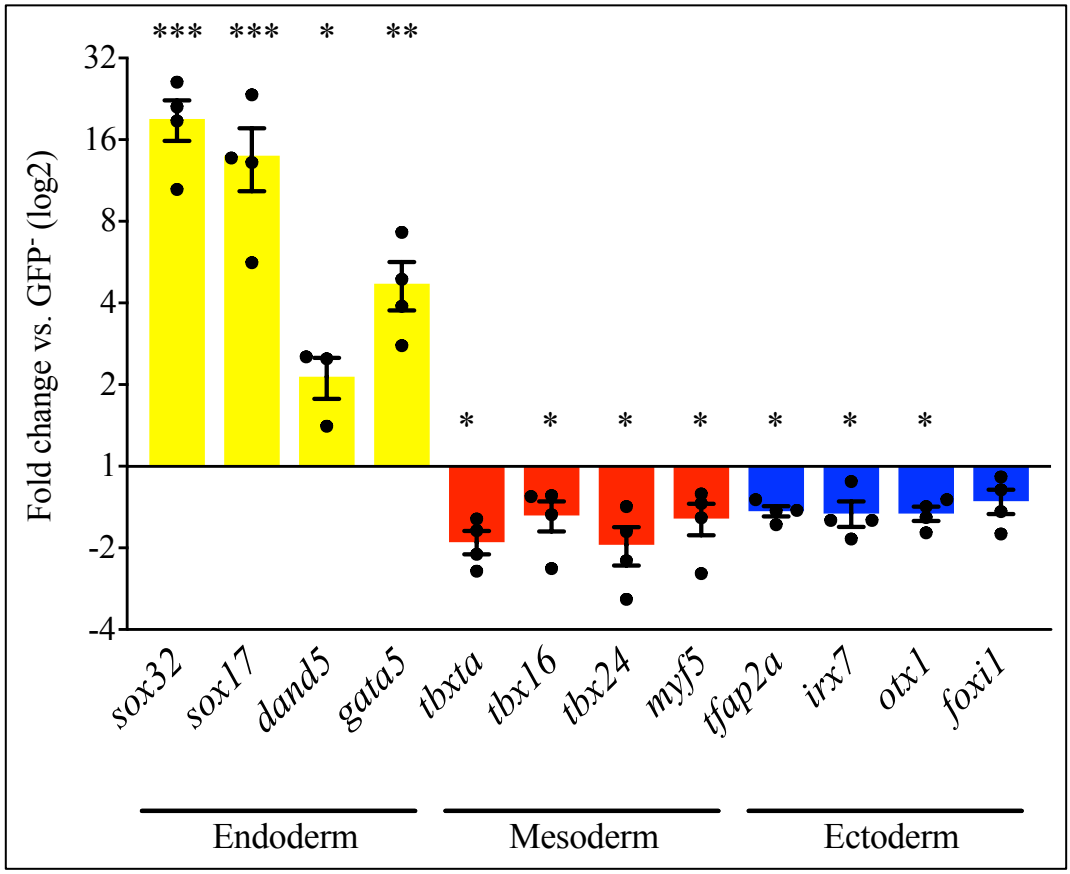


Figure 4.39 Markers of endoderm, mesoderm and ectoderm in non leaky embryos. Transcripts were quantified in the GFP$^-$ and GFP$^+$ expressing populations and are plotted as fold change relative to the GFP$^-$ population. Values > 1 indicate higher gene expression in GFP$^+$ cells, while values < 1 indicate

higher expression in GFP$^-$ cells. Fold changes are calculated in log2 base. Bars represent mean expression ± SEM (n=4 biological replicates). Statistical analysis was performed using two-tailed t test: *$p \leq 0.05$, **$p \leq 0.01$, ***$p \leq 0.001$.

In contrast, no significant enrichment of endodermal markers was observed in FAC sorted GFP$^+$ expressing cells from leaky embryos (Figure 4.40). Markers for all three lineages (endodermal, mesodermal and ectodermal) were expressed at similar levels in both GFP$^-$ and GFP$^+$ populations, except for *sox17* which was significantly enriched in GFP$^+$ ($p \leq 0.05$), suggesting *gfp* expressing cells were not exclusively endodermal.

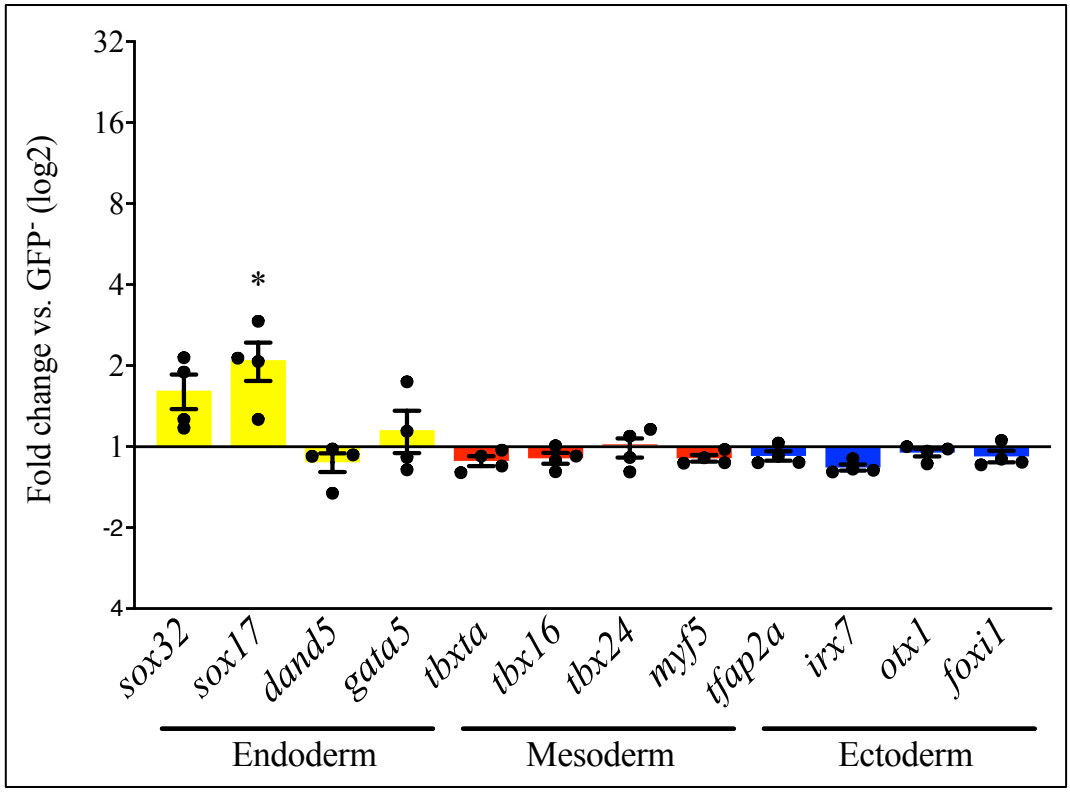


Figure 4.40 Markers of endoderm, mesoderm and ectoderm in leaky embryos. Transcripts were quantified in the GFP$^-$ and GFP$^+$ expressing populations and are plotted as fold change relative to the GFP$^-$ population. Values > 1 indicate higher gene expression in GFP$^+$ cells, while values lower < 1 indicate higher expression in GFP$^-$ cells. Fold changes are calculated in log2 base. Bars represent mean expression ± SEM (n=4 biological replicates). Statistical analysis was performed using two-tailed t test. *$p \leq 0.05$.

To further characterise the observed pattern, I proceeded to subdivide the GFP$^+$ population of cells from leaky embryos, whose GFP intensity ranged from 2 x 10^2 and 10^5, into two additional subpopulations I called 'high' and 'top' (Figure 4.40). Specifically, the window of

GFP intensity between 2 x1 0^2 and 10^3 became the 'high' population and the window between 10^3 and 10^4 became the 'top' population.

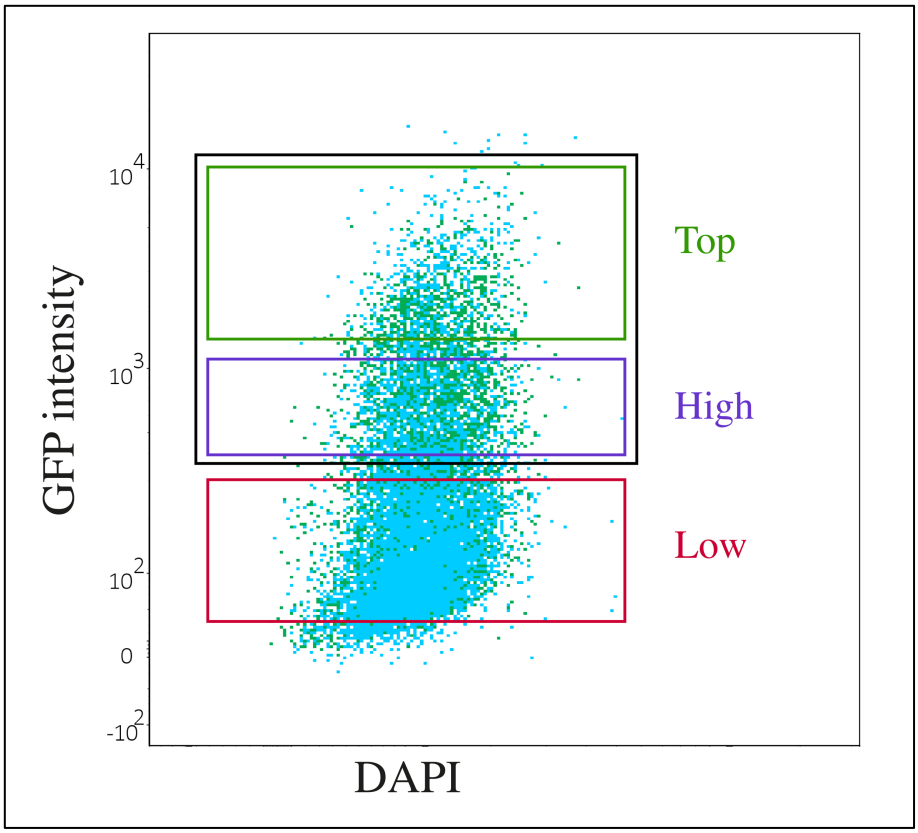


Figure 4.41 Additional subpopulation strategy to sort GFP$^+$ cells from leaky embryos. The previous high GFP/GFP$^+$ gate (black) was subdivided into 2 further internal population, 'high' (purple) and 'top' (green) to allow for more precise comparisons of gene expression in relation to level of GFP expression. The previously defined low GFP/GFP$^-$ gate is shown in red. Two biological replicates are shown in green and blue respectively.

When I used this methodology to compare gene expression profiles amongst the three different GFP intensity populations (low, high and top), three out of four endodermal genes (*sox32*, *sox17* and *gata5*) showed significant enrichment in the top population (Figure 4.42, yellow bars), whilst the other endodermal gene (*dand5*) showed no difference between cell populations. Of particular note, *sox32* and *sox17* transcripts levels were ~32- and ~38-fold enriched respectively, in the top population compared to the low population. In contrast, transcript numbers increased only ~2.6- and 4-fold respectively between the low and high populations. *dand5* fold change increased 3.1-fold in the high population but decreased again to 1.1-fold in top population, suggesting this change in expression was not linked to GFP intensity. Mesodermal and ectodermal genes were seen to be depleted in the top population,

particularly, *tbx16*, *tbx24* and *tfap2a* which showed ~1.5-, ~3.6- and ~1.9-fold reductions, respectively. Low populations and high populations of leaky embryos showed similar expression of mesodermal and ectodermal genes, suggesting that the transcript profile of these cells more closely resembled the transcript profile of cells with low GFP expression of non leaky embryos, and that the transcript profile of cells with top GFP expression (endoderm) of leaky embryos more closely resembled high GFP expression of non leaky embryos (compare Figure 4.42 to Figure 4.29). Ultimately, this suggested that only the brightest top GFP cells are truly endodermal and that the high GFP cells from leaky embryos, despite being GFP^{+}, are more representative of mesoderm and ectoderm.

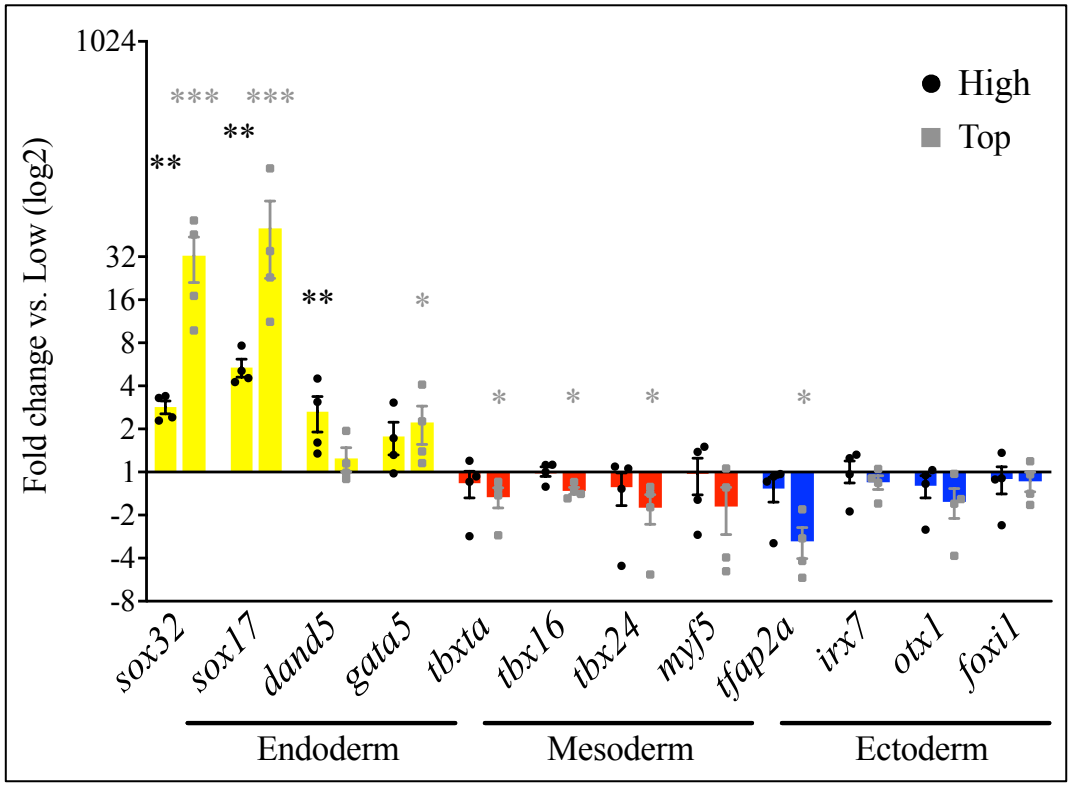


Figure 4.42 Markers of endoderm, mesoderm, ectoderm in high and top GFP expressing populations. Transcripts were quantified in low, high and top GFP populations and plotted as fold change relative to low population. Values >1 indicate higher gene expression than low GFP cells whilst values < 1 indicate higher expression values in low GFP cells. Fold changes are calculated in log2 base. Bars represent mean expression ± SEM (n=4 biological replicates). Statistical analysis was performed using one-way ANOVA: *$p \leq 0.05$, **$p \leq 0.01$, ***$p \leq 0.001$. Both *p-values* and points are colour coded: black High pollution, grey Top population.

From these data, I concluded that leaky embryos lacked the ability of WT and non leaky embryos to fine tune the expression of multiple genes and that they do not faithfully recapitulate gene dynamics during endoderm formation, at least during gastrulation. I therefore decided that leaky embryos were not suitable to study the dynamics of endodermal genes

during gastrulation. As such, I separated leaky embryos from non leaky embryos in every batch, and isolated GFP^+ cells only from non leaky embryos in order to prepare RNA-seq libraries as descried in Chapter 5.

4.7.5 Heterozygous and homozygous *sox17:GFP* embryos

I next sought to address whether there was a difference in GFP intensity between heterozygous and homozygous embryos. I hypothesised that, due to both alleles carrying the fluorophore, the GFP intensity would be higher in homozygous embryos. This was subsequently confirmed as GFP^+ cells from homozygous embryos showed a higher GFP intensity and were more readily detected during FACS compared with heterozygous siblings. In addition, more GFP^+ cells were recovered from dissociating the same starting number of homozygous *sox17:GFP* embryos compared to heterozygous siblings (Figure 4.43).

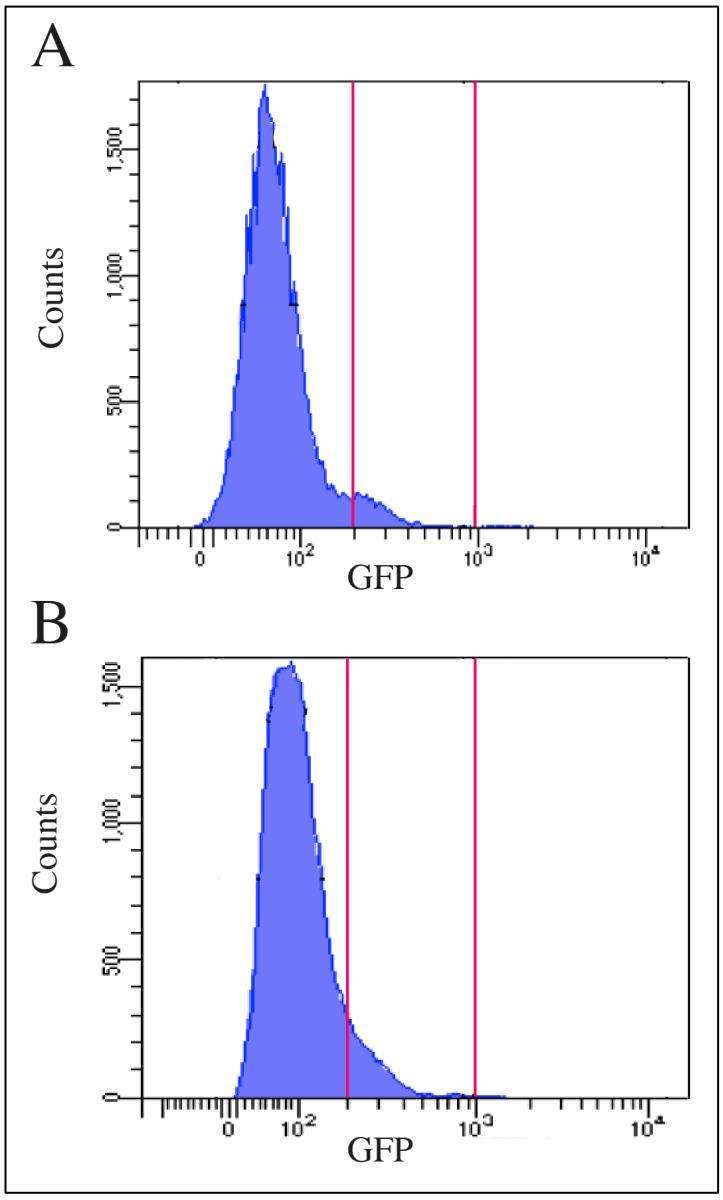


Figure 4.43 Representative example of sorted embryos from homozygous and heterozygous *sox17:GFP* embryos. (**A**) Heterozygous cell populations (50 embryos) and (**B**) homozygous cell populations (50 embryos) subject to FACS were compared.

For the homozygous *sox17:GFP* sorted cells, the proportion of GFP^+ cells in the sample was 24 ± 6% in all 4 batches analysed. For the heterozygous *sox17:GFP* sorted cells, this ranged from 12.15 to 29%, with a mean of 19 ± 6%. This range was comparable to that of sorted homozygous cells, however the MFI diverged between the two conditions. Homozygous embryo derived cells yielded MFI of 642. In contrast, heterozygous derived cells yielded MFI of 584. The recovered percentage of endodermal cells from homozygous embryos was consistent with the numbers of endodermal cells retrieved in two independent published

scRNA-seq studies (Farrell et al., 2018; Wagner et al., 2018), where endodermal cells comprised 30% of all cells at 9.00 hpf.

Due to the higher number of cells recovered from embryos from homozygous *sox17:GFP* adults, less time spent sorting (leaky embryos were easier to spot and faster to remove than GFP^- embryos) and the similar ratio of GFP^+ cells to GFP^- cells amongst the two genotypes, I proceeded to use only homozygous embryos for all the follow up experiments.

4.8 Technical issues and methodology justification

To isolate endodermal cells, I used the transgenic *sox17:GFP* line that expresses the fluorophore GFP in endodermal precursors during gastrulation. RT-qPCR for specific known endoderm, mesoderm and ectoderm markers was performed on cDNA synthesized from FAC-sorted GFP^+ cells to ensure that I was enriching only for the desired endodermal cell population and excluding cells of other lineages. Within my isolated putative endodermal cell populations, I then compared gene expression levels of known endodermal, mesodermal and ectodermal markers between the GFP^+ cell population and GFP^- cell population.

Different methods of quantifying mRNA levels have been established using multiple platforms. For my experiments, I first applied dye-based chemistry methods (SYBR green) to quantify transcript abundance, before switching to probe-based chemistry and multiplex qPCR solutions for mRNA quantification. This is because the latter proved to be more accurate than the standard SYBR green method and was therefore a more precise way of quantifying gene enrichment in my GFP^+ sorted populations.

In the first instance, I simultaneously profiled the expression levels of 12 different genes (endodermal, mesodermal and ectodermal) in three biological replicates from sorted cells, using the SYBR green approach. Ct values of the target genes in both GFP^+ and GFP^- cells were normalised to the reference gene (*actb*) and plotted as fold change relative to the expression levels in GFP^- (Figure 4.44). The GFP^+ cells showed enrichment for the endodermal markers *sox32* (11-fold change) and *sox17* (17-fold change) whereas, *gata5* and *dand5*, two other endodermal markers were not seen to be significantly enriched in the GFP^+ cells. In addition, expression levels of markers of mesoderm and ectoderm were similar in both the GFP^- and GFP^+ populations and I therefore did not observe the expression pattern I was expecting: downregulation of mesodermal and ectodermal genes in the GFP^- population. I was

therefore not convinced of the enrichment of a pure endodermal cell population based only on the *sox17* and *sox32* endodermal markers showing a significant positive fold change. I also observed that mesodermal and ectodermal gene expression levels were under enriched in the GFP$^-$ population, again differing from predictions that these cells were more likely to be of meso or ectodermal character.

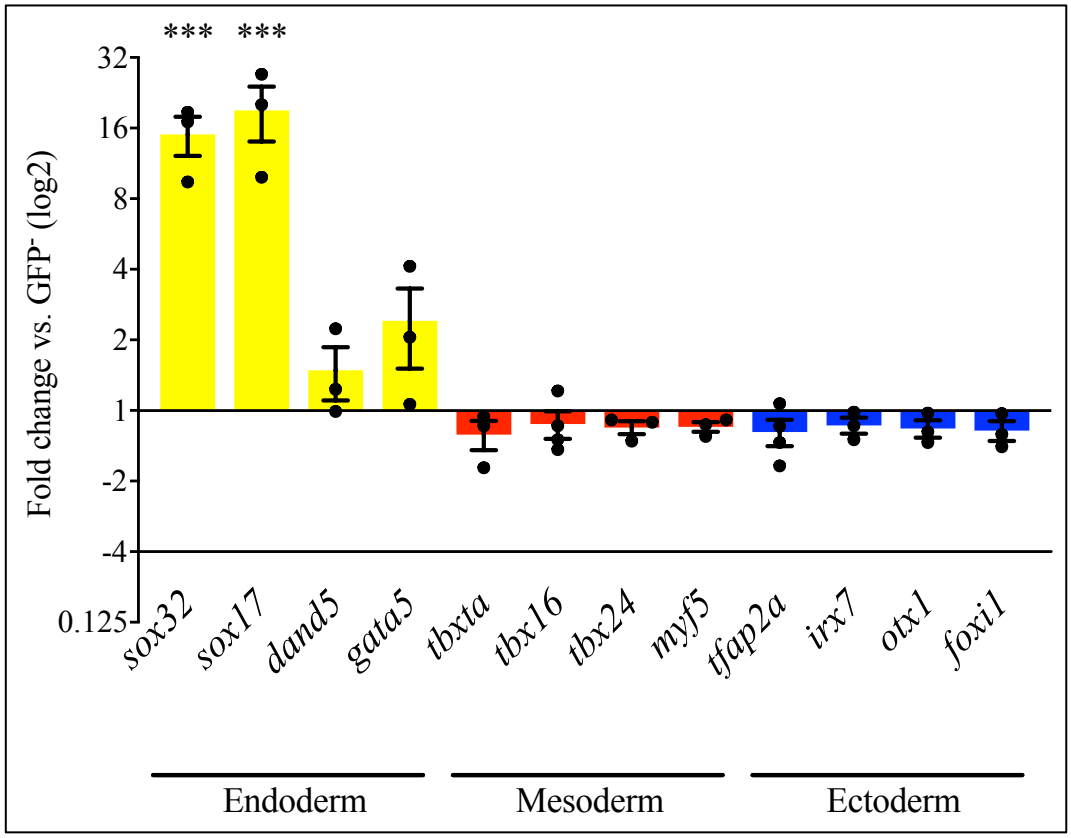


Figure 4.44 Fold change comparisons of 12 target genes relative to expression in GFP$^-$ cells using SYBR-green RT-qPCR. Endodermal (yellow bars), mesodermal (red bars) and ectodermal (blue bars) genes are shown. Note the positive enrichment for *sox32* and *sox17* endodermal genes in GFP$^+$ cells compared to GFP$^-$ cells. No difference in fold change for any other endodermal genes, nor any mesodermal and ectodermal genes was observed. Data is shows as mean + SEM. Statistical analysis was performed using Student's t-test (two-tailed) (n =3). ***$p \leq 0.001$.

I therefore decided to try probe-based quantitative PCR (TaqMan chemistry). This method offers the least background fluorescence as compared to other dye-based chemistries, combined with high sensitivity that can detect a single transcript copy and is therefore recommended for use when measuring transcript abundance in low RNA content samples (Wong et al., 2015). Crucially, probe-based assays allow multiplexing, meaning that in a single reaction tube multiple targets can be amplified by different sets of primers, with a unique reporter probes that distinguish each PCR amplicon. Thus, the expression levels of several

genes of interest are measured simultaneously, minimizing the amount of starting material required. This was of critical value as the sorted sample size was limited by the number of embryos obtained from the spawning fish clutch.

Probe-based qPCR is based on the detecting hydrolysis of the fluorescently-tagged probe. Each probe has a fluorescent reporter and a quencher molecule that prevents light emission. These two species are in sufficiently close proximity to prevent any fluorescence emission. However, during the elongation and extension phases of the PCR cycle, the exonuclease activity of the polymerase hydrolyses the probe, freeing the reporter from the quencher molecule and fluorescent light is emitted. The amount of PCR product generated is directly proportional to the increase in fluorescence, allowing a more accurate quantification of the amplified target than dye-based chemistry allows, where the dye is intercalated into the double helix of DNA molecules.

To demonstrate the utility of multiplex PCR, I first showed that gene expression in a multiplex experiment provided similar results to singleplex experiments, before performing any sample enrichment comparisons in my sorted cell populations. I performed an initial profiling experiment where each gene assay was run both individually and together in a multiplex reaction. Reactions were performed on the same instrument (Abi ViiA7) using the same input material and the resulting transcript levels are illustrated in Figure 4.45. High concordance between the Ct values of both the singleplex and multiplex reactions were clear. No significance differences were observed between the reaction types meaning the multiplex reagents afforded similar efficiencies to the singleplex reactions using the same probes.

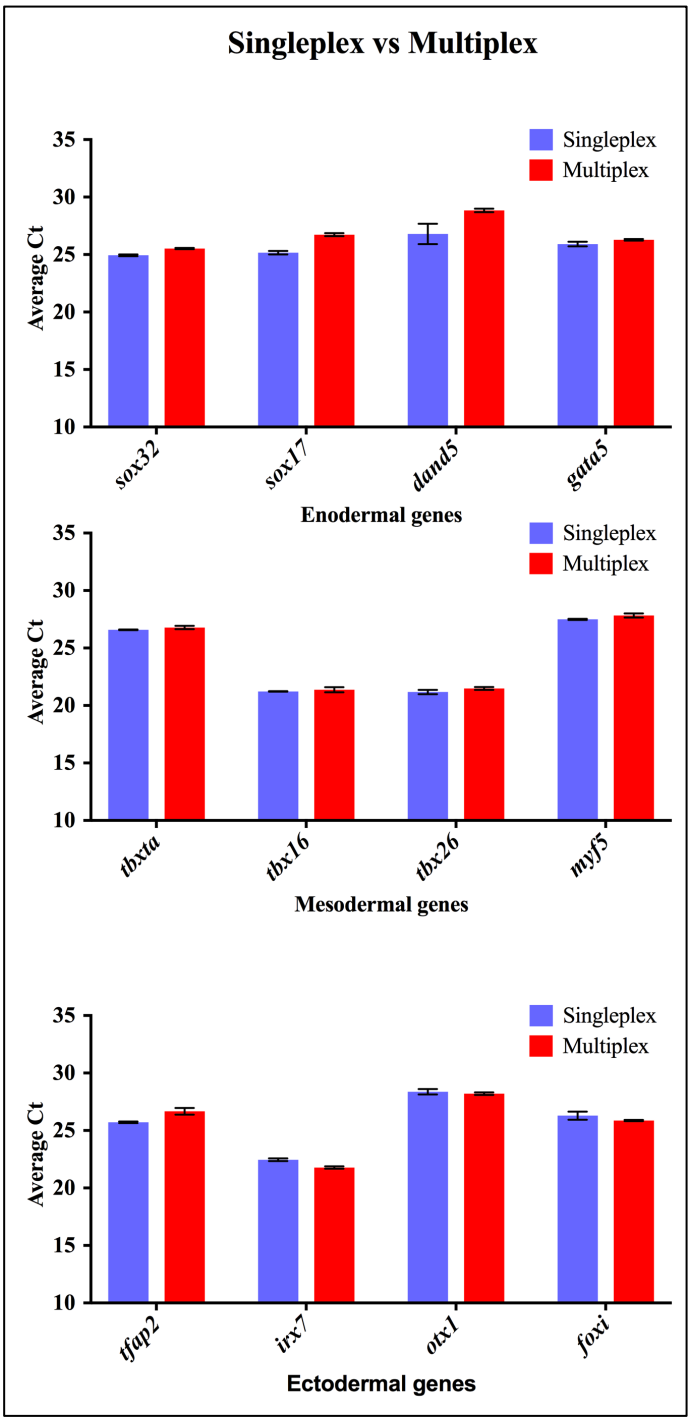


Figure 4.45 Singleplex vs multiplex RT-qPCR. Comparison of average Ct values obtained from singleplex (purple bars) and multiplex (red bars) RT-qPCR reactions for twelve genes. cDNA from 9.00 hpf WT embryos was used. Data are shows as mean + SEM (n=3). No statistically significant differences were observed between the 2 reaction types (Mann-Whitney U test).

I then set up a series of cDNA concentrations to determine the minimal amount of starting material required where multiplex and singleplex reactions would remain in concordance and the Ct values would not be artificially reduced by the depletion of reagents in the multiplex reaction. For all the probes used, highly significant linear curves between the amount of

starting DNA and the Ct values were obtained from 25 ng to 40 pg, two examples of which are shown in Figure 4.46.

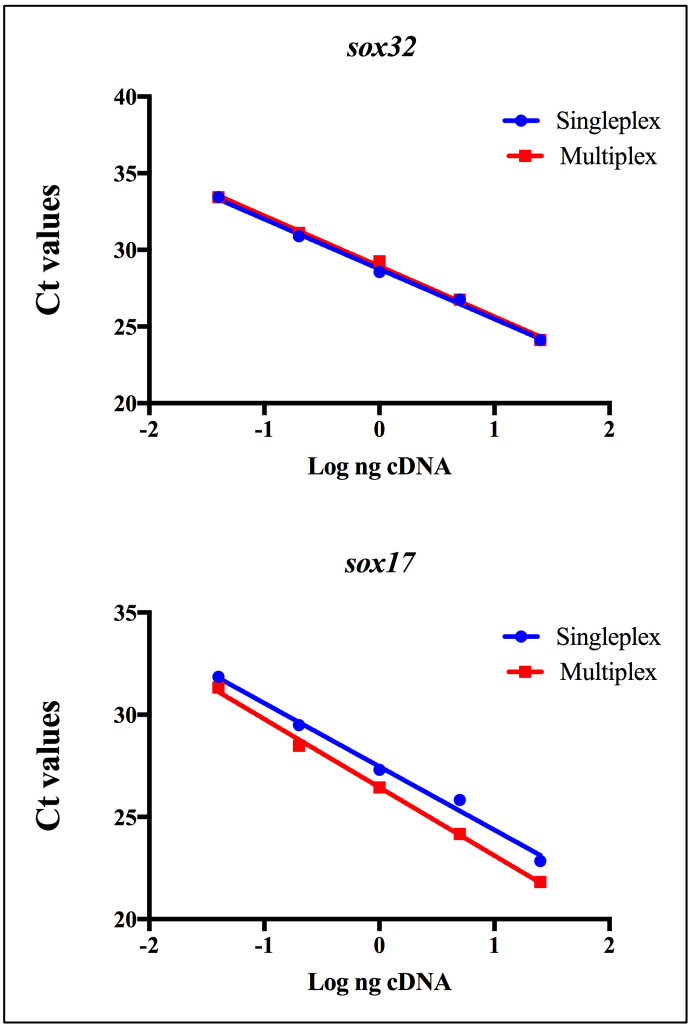


Figure 4.46 Examples of linear range of singleplex and multiplex reaction in relation to cDNA starting amount. *sox32* (top) and *sox17* (bottom) primers were used to analyse the relationship between starting amount of cDNA and Ct values. Amplification was performed using a serial dilution of cDNA from ranging from 25 pg to 40 pg per 10 µl reaction. Singleplex (blue) and multiplex (red) are shown.

For the *sox32* and *sox17* reactions shown in Figure 4.46, the regression equations were: $y = -3.280 \log(x) + 5.536$ ($R^2 = 0.996$, n =3) and $y = -3.101 \log(x) + 8.728$ ($R^2 = 0.997$, n =3) respectively. The PCR efficiency and coefficient of correlation coefficient (R^2) for all 12 genes (plus the *actb* reference gene) are summarized in Table 3.

Table 4.3 PCR efficiencies and R^2 values for all genes studied in both singleplex and multiplex reactions.

	PCR efficiency (%)		Coefficient of correlation (R^2)	
Gene	**Singleplex**	**Multiplex**	**Singleplex**	**Multiplex**
actb	99.2	99.5	0.997	0.984

sox32	100.9	100.7	0.993	0.996
sox17	99.6	105.1	0.991	0.997
dand5	102.6	103.3	0.986	0.987
gata5	103.5	105.6	0.995	0.944
Ntl	104.8	106.9	0.996	0.981
tbx16	96.7	94.0	0.993	0.993
tbx26	105.8	105.0	0.987	0.987
myf5	100.7	100.6	0.981	0.982
tfap2	99.7	95.8	0.989	0.963
irx7	103.9	102.6	0.989	0.989
otx1	102.7	103.2	0.995	0.995
foxi1	102.6	101.5	0.988	0.982

From these data, I chose 1 ng as the input amount to be used in all assays because it showed less average difference between the singleplex and multiplex reactions.

I next evaluated the ability of multiplex RT-qPCR to analyse mRNA levels in sorted cell populations and compared transcript levels between GFP^- and GFP^+ populations and calculated the average Ct values (n=3 technical replicates) for each assay in both single and multiplex reactions in 2 biological replicates (Figure 4.47). Notably, the 13 singleplex reactions (12 targets genes + reference gene, in triplicate) required a total of 39 ng of cDNA while the multiplex reactions needed only 13 ng. This difference was considered extremely important due to the low RNA yield of sorted cells. As predicted, I observed a high correlation in values between the singleplex and multiplex measurements using 1 ng cDNA starting material and the same trends between the two different measurement strategies when comparing enrichment between GFP^- and GFP^+ populations (Figure 4.47). However, it should be noted that for *sox32*, *dand5* and *tfap2a,* the signal intensity was statistically significantly different between the singleplex and multiplex reactions suggesting that singleplex qPCR was more sensitive than multiplex qPCR in these instances.

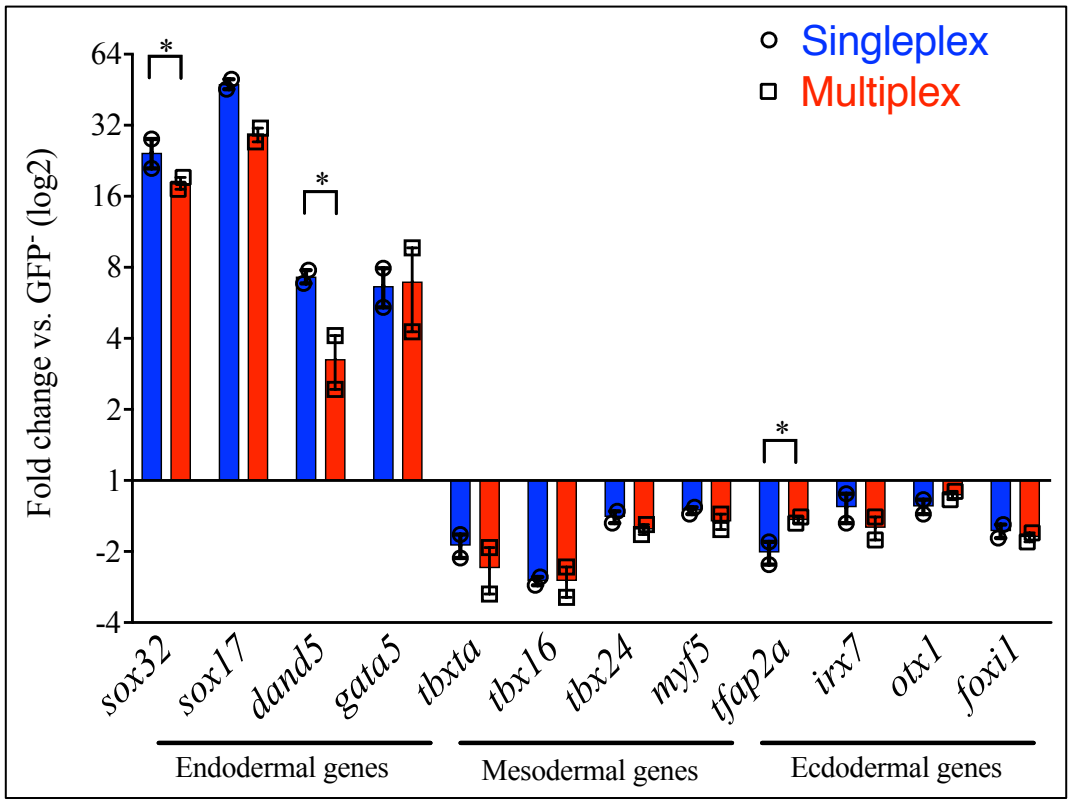


Figure 4.47 Fold change comparison of 12 targets to GFP$^-$ cell gene expression levels using singleplex and multiplex RT-qPCR format. The differences in transcript abundance measured by multiplex RT-qPCR were identical to traditional singleplex RT-qPCR with the exception of *sox32*, *dand5* and *tfap2a* where a statistically significant difference is observed. Each point is a biological replicate, each being the mean of a technical triplicate. Statistical analysis was performed using Students t-test (two-tailed) (n=2). *$p \leq 0.05$.

All endodermal genes were seen to be upregulated (e.g. multiplex: *sox32* 17-fold; *sox17* 31-fold; *dand5* 3-fold; *gata5* 7-fold) while mesodermal and ectodermal genes were downregulated (e.g. multiplex: *tbxta* -2-fold; *tbx16* -2.4-fold; *tbx24* -1.8-fold) and all were in concordance with the singleplex results, with the exceptions outlined above.

I then sought to compare the differences in sensitivity (the ability to detect transcripts) between TaqMan chemistry and SYBR-green. To do so, I compared fold change values achieved using the same input amount for both chemistries. The results from both strategies generally matched, however a noticeable wider spread of SEM was observable for SYBR samples, which indicated higher technical variability in the samples, potentially including loading and pipetting errors. Importantly, this variability caused the loss of statistical significance in the levels of downregulation observed in five mesodermal and ectodermal genes between GFP$^-$ and GFP$^+$ populations (Figure 4.48, blue squares). All fold change values for downregulated genes were found to be lower in the SYBR samples and significantly

different from the TaqMan samples for *tbxta*, *tbx16* and *irx7* ($p \leq 0.05$). *foxi* and *otx2* were borderline significantly different between the two chemistries ($p = 0.07$). Most importantly, the two singleplex reactions required a total of 65 ng of cDNA input, whereas the multiplex assay reaction required only 26 ng. This clearly illustrates how multiplex PCR is capable of extracting more information from a smaller amount of starting sample.

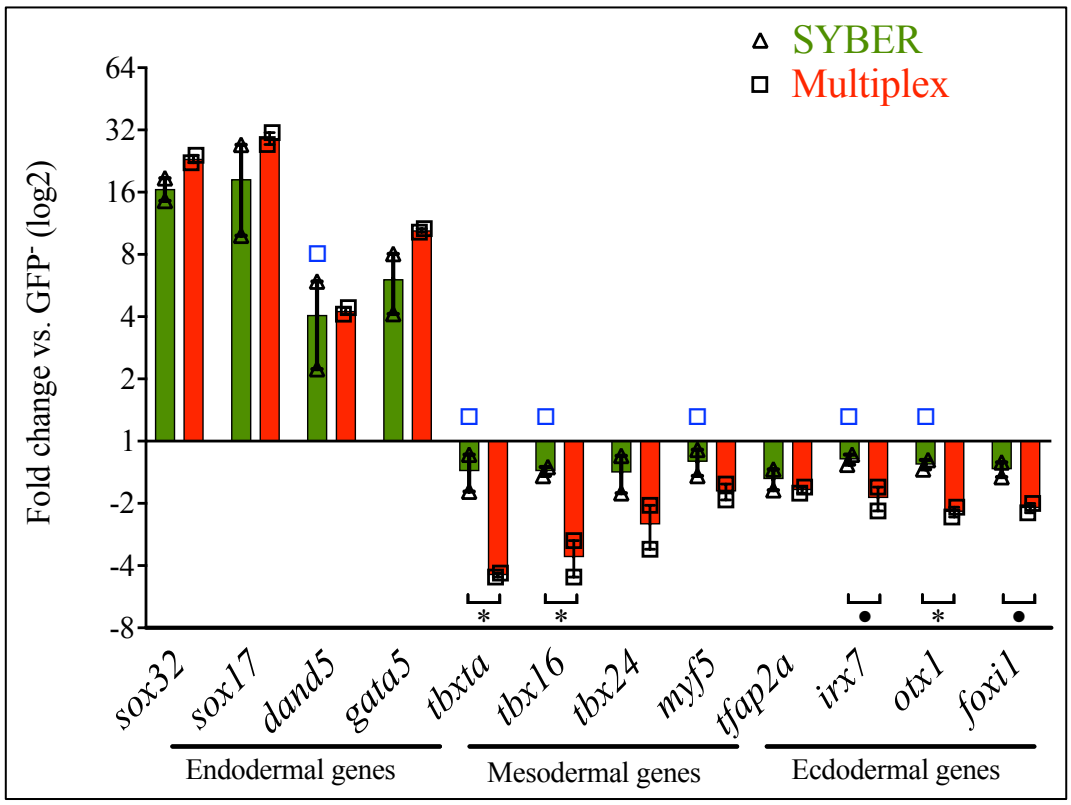


Figure 4.48 Comparison of SYBR and TaqMan multiplex gene expression profiles. Fold changes in gene expression level relative to the level in GFP$^-$ expressing cells are shown for the endodermal, mesodermal and ectodermal genes tested. TaqMan multiplex RT-qPCR (red bars) was the more sensitive chemistry in detecting mRNA levels compared to SYBR (green bars). Blue squares denote genes where the loss of statistical significance occurred between GFP$^-$ and GFP$^+$ populations. Statistical analysis was performed using Student t-test (two-tailed) (n=2) *$p \leq 0.05$; •$p = 0.07$.

Taken together, these data showed that the TaqMan probes were more suitable to use in the quantification of gene expression in cell populations enriched for putative endodermal precursors. Compared to the traditional SYBR method, TaqMan multiplex RT-qPCR showed better specificity, a higher amplicon efficiency and crucially, minimized the amount of starting material required because multiple targets were amplified in a single reaction tube. This latter advantage was of critical value when analysing the RNA extracted from sorted cells, as the starting material available was a key limiting factor. I concluded that the TaqMan multiplex

method was reproducible and specific, and could detect, with accuracy, the limited number of transcripts present in my sorted cell samples. I therefore applied multiplex RT-qPCR first to the sorted samples to assess the differences in gene expression between leaky and leaky embryos and then as a crucial control for the input samples used for my RNA-seq experiments (described in Chapter 5).

4.9 Chapter Summary

This chapter sought to characterise the transgenic *sox17:gfp* reporter line and isolate a population of endodermal cells at the end of gastrulation, in doing so, I investigated variations in GFP expression at both the mRNA and protein level, and changes in endogenous gene expression associated with these variations. Studies using genetically encoded reporter genes such as GFP have provided valuable information about complex cellular processes, and detailed studies of the dynamics of spatial and temporal gene expression are made possible by genetically encoding a fluorophore under the control of various promoters (Gong et al., 2001). However, using this *sox17:GFP* transgenic line, I noticed embryos from the same clutch that showed higher GFP fluorescence when observed under the microscope. I followed up this observation and showed that what I saw in live embryos was supported by immunohistochemistry data; higher fluorescence intensity was related to a higher abundance of the protein in what I then called leaky embryos. I then proceeded to collect zebrafish at different stages of development; interestingly, the percentage of leaky offspring from the founders ranged from 3% to 13%, which was not linked to maternal-effect regulation in activating the transgenic locus and the nature of leakiness was stochastically mosaic. I proceeded to use the embryos I had collected to evaluate the expression levels of selected genes marking the three germ layer lineages, under the hypothesis that a higher level of GFP expression could mean that the gene regulatory system was being perturbed. In order to do so, I tested a series of different gene expression primers to interrogate whether the leaky embryos showed a different expression pattern compared to the non leaky embryos, and how both relate to expression levels in WT embryos. Although the higher distribution of GFP protein matched that of *gfp* mRNA during gastrulation, by 24 hpf, a stark contrast was seen between mRNA and protein levels; the level of GFP protein in leaky embryos continued to be significantly higher in leaky embryos, but by this time point, there was no difference in *gfp* mRNA expression between non leaky and leaky embryos. Moreover, during gastrulation, leaky embryos exhibited significantly higher expression levels of genes associated with endoderm

(*sox32* and *sox17*), misexpression of early mesendoderm (*crcx4* and *mixl1*) and lower expression of genes associated with axial and paraxial mesoderm (*tbxta* and *myf5* respectively). By the 24 hpf timepoint, these expression levels had reverted back to those consistent with non leaky and WT embryos.

As previously discussed, given the dynamics of GFP production (the fluorophore takes some time to form after mRNA synthesis) and its relative stability (Corish and Tyler-Smith, 1999), it is likely the disparity observed between mRNA and protein expression levels at 24 hpf can be explained by the half life of GFP protein. mRNA levels of *gfp* were already identical in leaky and non leaky embryos after 24 hours, while the cells positive for GFP were still significantly enriched in leaky embryos at the same time. This suggests that the disparity in the regulatory mechanism that caused GFP expression between leaky and non leaky fish was no longer in place, causing the promoter to be turned off in non endodermal cells at this point. Therefore, GFP is by now only produced in endodermal cells, and the observed difference disappeared once the remaining ectopically expressed GFP had been degraded to background levels. This stage has already been reached by 48 hpf, by that time no difference in GFP transcript or protein levels could be observed between the groups. However, I am not currently able to propose a mechanism to explain the temporal nature of the misregulation of the affected genes, and more detailed studies into the dynamics of the spatial and temporal onset and cessation of these genes is required. Furthermore, no explanation regarding the absence of phenotypic changes in leaky embryos is proposed. However, the data described in this chapter are robust enough to allow me to speculate that changes in gene expression in leaky embryos may well occur during the specification of endoderm in leaky embryos, and it is wholly possible that changes in the expression levels of these genes may trigger some form of compensatory mechanism. The early embryo needs to be, and indeed is, a remarkably robust, resilient and adaptable biological structure (Macneil and Walhout, 2011; Osterwalder et al., 2018; Sharifi-Zarchi et al., 2015; Wagner, 2008). More recently, genetic compensation mechanisms in response to DNA lesions have been described in the zebrafish embryo, specifically following the use of CRISPR technology (Rossi et al., 2017). Furthermore, the phenomenon of transcriptional adaptation has been described, whereby compensatory mechanisms are employed by the embryo following detection of aberrant changes at the level of the mRNA (El-Brolosy et al., 2018; Smith et al., 2013). Further studies are warranted to determine the mechanisms at play in leaky embryos, that seemingly prevent the observed aberrant gene expression levels from having any apparent deleterious effects, in particular how

higher numbers of endodermal cells and lower numbers of mesodermal cells during the gastrulation process harbour no visible defects in endo-mesodermal structure such as liver, pancreas and kidneys.

Whilst it is useful to look at the endoderm specification pathway as a linear cascade of genes that are chronologically induced and then repressed, the reality is much more complex, with intricate networks of TFs intersecting to form multiple interacting pathways. It is therefore reasonable to postulate that the changes in gene expression that occur in leaky embryos during gastrulation may well be normalised by compensatory gene regulatory mechanisms that result in no phenotypical or genetic abnormalities by the end of somitogenesis and, as mentioned above are tightly linked to the evolution of mechanisms of functional redundancy in biological systems .

Chapter 5 – RNA-seq on endodermal related zebrafish line

Chapter 5 highlights:

- Identification of differentially expressed genes during zebrafish endoderm development using RNA sequencing (RNA-seq) on different fish lines through:
  - Bulk mRNA-seq of *sox32*$^{-/-}$ mutants at 5.25 hpf and 9.00 hpf
  - Bulk RNA-seq of *mixl1*$^{-/-}$ mutants at 5.25 hpf
  - Bulk mRNA-seq for FAC-sorted endoderm cells from *tg(sox17:GFP)* embryos 9.00 hpf
- Presentation of a streamlined bioinformatics pipeline for data analysis of RNA-seq.
- Identification and validation of new endodermal markers.

5.1 Introduction

How do developing zebrafish embryos control both the spatial and temporal expression of developmental genes? The coordination of multiple TFs is needed to orchestrate the formation of the 3 primary germ layers. Many studies have focused on defining the critical TFs that specify tissue identity and control morphogenesis and various fate maps have been created over the years which have helped shed light on how so-called 'master regulator' genes operate in a hierarchy of gene expression to characterise the multiple cell lineages within the embryo (Chan and Kyba, 2013; Davis and Rebay, 2017; Mattick et al., 2010). In zebrafish, studies have identified how Sox32, in combination with other specific partners such as Sox17 and Mixl1, allows the transcriptional activation of a set of endoderm-specific markers, however no studies have investigated the downstream set of genes regulated by these 3 TFs or captured gene expression information on a purified endodermal cell population.

A great deal of information on endoderm biology was classically obtained by perturbation of development systems; in particular, classic gain- and loss-of-function approaches using RNA injections and morpholino oligonucleotides (MO) respectively (Erter et al., 1998; Kikuchi et al., 2000; Poulain et al., 2006; Reiter et al., 1999; Rodaway et al., 1999; Schier et al., 1997; Warga and Nusslein-Volhard, 1999). Observing the phenotypes associated with knockdown of specific genes causing malformations or defects in comparison to wildtype

embryos, has provided us with information on the roles of specific endodermal genes. For example when Bjornson et al. (2005) injected a *mixl1* MO together with an *eomes* MO followed by the observation of a significant reduction in endoderm, suggesting that *eomes* and *mixl1* combinatorically contribute to specify endodermal lineages. Similar data have helped us to assemble a basic endodermal differentiation pathway and to understand more about the fate decisions cells make as they progress through development. However, in order to understand the complete picture, the study of individual genes is not enough; a genome-wide search for *sox32 and mixl1* downstream target genes and other endoderm-specific transcription factors is important to identify key regulatory inputs (direct or indirect) into the regulation of many of these genes and expand the analysis of the endodermal GRN by reconstructing programs of differential gene expression in the endodermal dynamic networks of regulatory genes. *Sox32* has arisen in zebrafish as a result of genome duplication and has evolved in a way not found in other species. The overall role of *sox17* in other species seems to be matched to the combined action of *sox32* and *sox17* in zebrafish. Complementary analysis of *mixl1,* which is required for *sox32* activation in endoderm formation, are also required to provide a complete genome wide picture of the functions of these key endodermal genes. By carrying out an RNA-seq analysis of these 3 effectors (*sox32*, *sox17* and *mixl1*) and by identifying their key regulatory outputs, the developmental functions of these key genes can be better understood. The regulatory control of several genes downstream of these 3 effectors was a key step in the careful construction and curation of the network that links the cell's genome and gene expression to endodermal cell identity.

In the last decade, next-generation sequencing (NGS) technologies have been widely used in the life sciences, and RNA-seq has become the most widespread method for analysis of the transcriptome; revealing the presence (or absence) and quantity of RNA in a biological sample under specific conditions (Ari and Arikan, 2016; McGettigan, 2013; Qian et al., 2014; Zhang et al., 2011).

RNA-seq technology was developed to offer a more comprehensive understanding of the transcriptome and overcome the limitations of cDNA microarrays which rely on existing knowledge about genome sequence, have high background levels , a limited dynamic range of detection due to signal saturation and need complicated normalisation methods to compare different experiments. RNA-seq was also shown to detect lowly expressed transcripts with reduced false positive rates in comparison to microarray based expression quantification

(Illumina, 2011; Nelson and Hurd, 2009; Zhao et al., 2014a; Zhao et al., 2014b). Since RNA-seq does not rely on a prespecified selection of cDNA probes, there are numerous additional applications of the technique that go beyond quantification of expressed transcripts of known genes, such as the detection and quantification of splice isoforms, fusion genes, novel transcripts and protein-RNA interaction sites. In addition, RNA-seq provides significantly more advantages due to its statistical power (Zhao et al., 2014a). In fish, RNA-seq has been applied to several species including zebrafish, channel catfish, European sea bass, rainbow trout, and grass carp to study numerous biological processes such as stress response, disease conditions and adaptive evolution (Gaither et al., 2018; Gurgul et al., 2018; He et al., 2017; Sarropoulou et al., 2012; Smith et al., 2013; Zeng et al., 2016), although the detection of gene expression changes (quantification of mRNA levels) between different cell populations and/or experimental conditions (WT vs mutant) remains the most common application of RNA-seq approaches in zebrafish (Hostelley et al., 2017).

In this chapter, I provide RNA-seq datasets on 2 different zebrafish mutants where the organisation of endoderm is perturbed, in order to better understand the transcriptional programs that lead to mesendodermal cell fate. These comprise the Mix paired-like homeobox factor (*mixl1*$^{-/-}$) and the Sry-related HMG box 32 (*sox32*$^{-/-}$) mutant lines. Transcriptomic analyses were performed on samples at 5.25 hpf for both mutants, with an additional time point of 9.00 hpf for *sox32*$^{-/-}$. Mixl1 expression terminates soon after the start of gastrulation, I therefore prioritised understanding the role of Mixl1 in the process of endoderm and mesoderm induction at 5.25 hpf by sequencing deeper and with higher coverage at this stage, rather than adding experimental replicates at 9.00 hpf. Furthermore, I used fluorescence-activated cell sorting (FACS) at 9.00 hpf to isolate and then transcriptionally profile endodermal cells, using the transgenic *sox17:GFP* reporter line described in Chapter 4. The same process at 5.25 hpf was unsuccessful due to technical limitations (low number of GFP^{+} cells). A total of 36 samples were analysed for differential gene expression, leading to the identification of putative new markers associated with endoderm formation. The overlap among these datasets helped identify new endodermal genes and detect and measure the extent of genetic overlap between *mixl1*, *sox32* and *sox17,* outlining a common biological pathway in endoderm formation. The data I generated in this study have been integrated together with recently published scRNA-seq datasets (Farrell et al., 2018; Wagner et al., 2018) to provide insights into the unique transcriptional program that leads to endoderm specification in zebrafish development.

5.2 Overview of RNA-seq workflow

Similar to all available genome-based high-throughput sequencing approaches, the RNA-seq field still lacks extensively accepted and adopted standards. More recently, ENCODE and similar consortiums, such as modENCODE and DANIO-CODE, have started to release general best practice guidelines for analysing big data and descriptive suggestions on how to analyse differential gene expression (Byron et al., 2016; Conesa et al., 2016; Dunham et al., 2012; Gerstein et al., 2010; Roy et al., 2010; Tan et al., 2016).

The general RNA-seq workflow can be separated into 2 phases, the first phase is wet-lab extraction of RNA and library preparation; the second phase is computational, where the sequenced samples are analysed. The RNA-seq workflow is summarised in Figures 5.1 and 5.2, for the wet-lab and computational phases respectively. Full details of the methodologies used can be found in Materials and Methods.

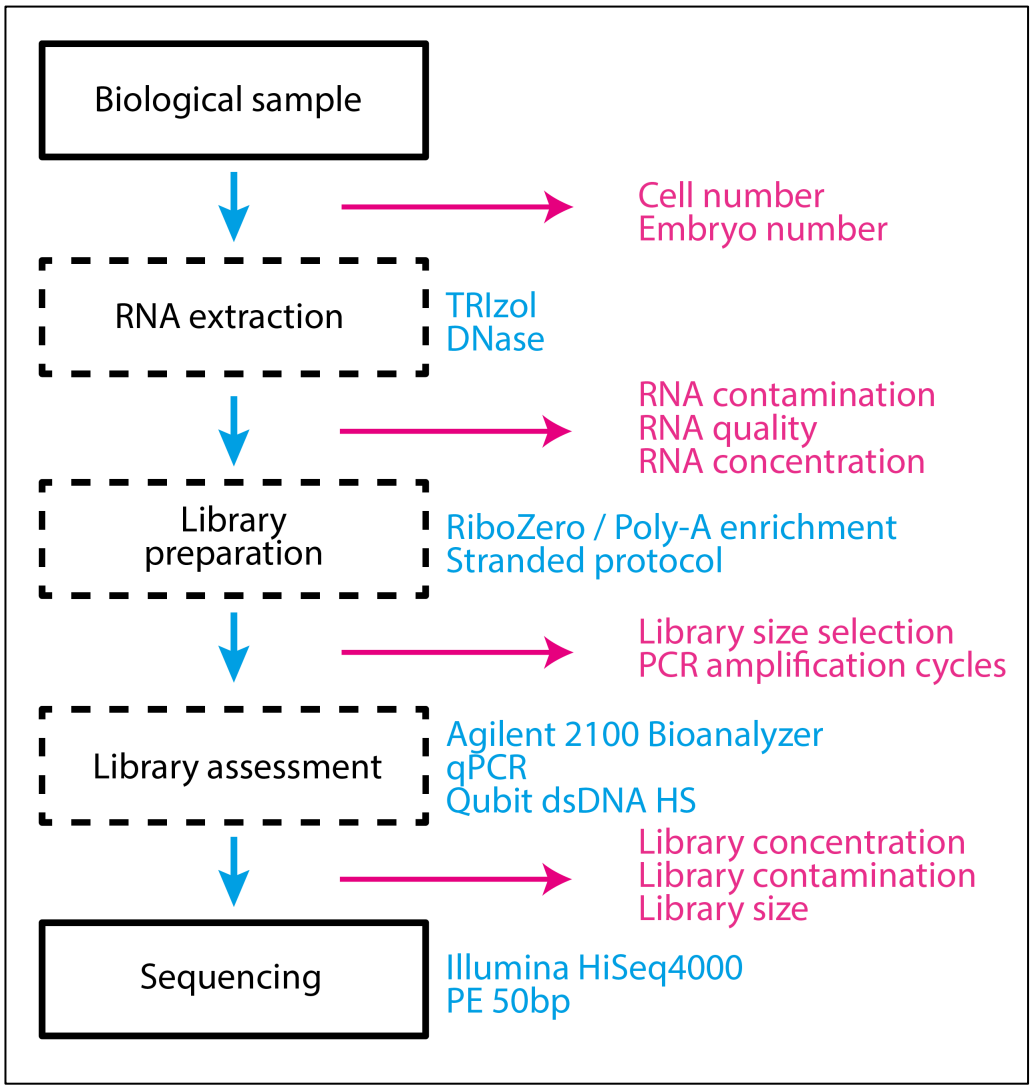


Figure 5.1 Workflow to extract and sequence RNA (wet-lab phase). Schematic representation of the workflow from obtaining the biological samples (cells/embryos) to sequencing the prepared libraries. Solid boxes denote start and end points; dashed boxes intermediate steps. Blue text denotes methodologies used for the respective step; pink text denotes quality control steps (see Materials and Methods for full details).

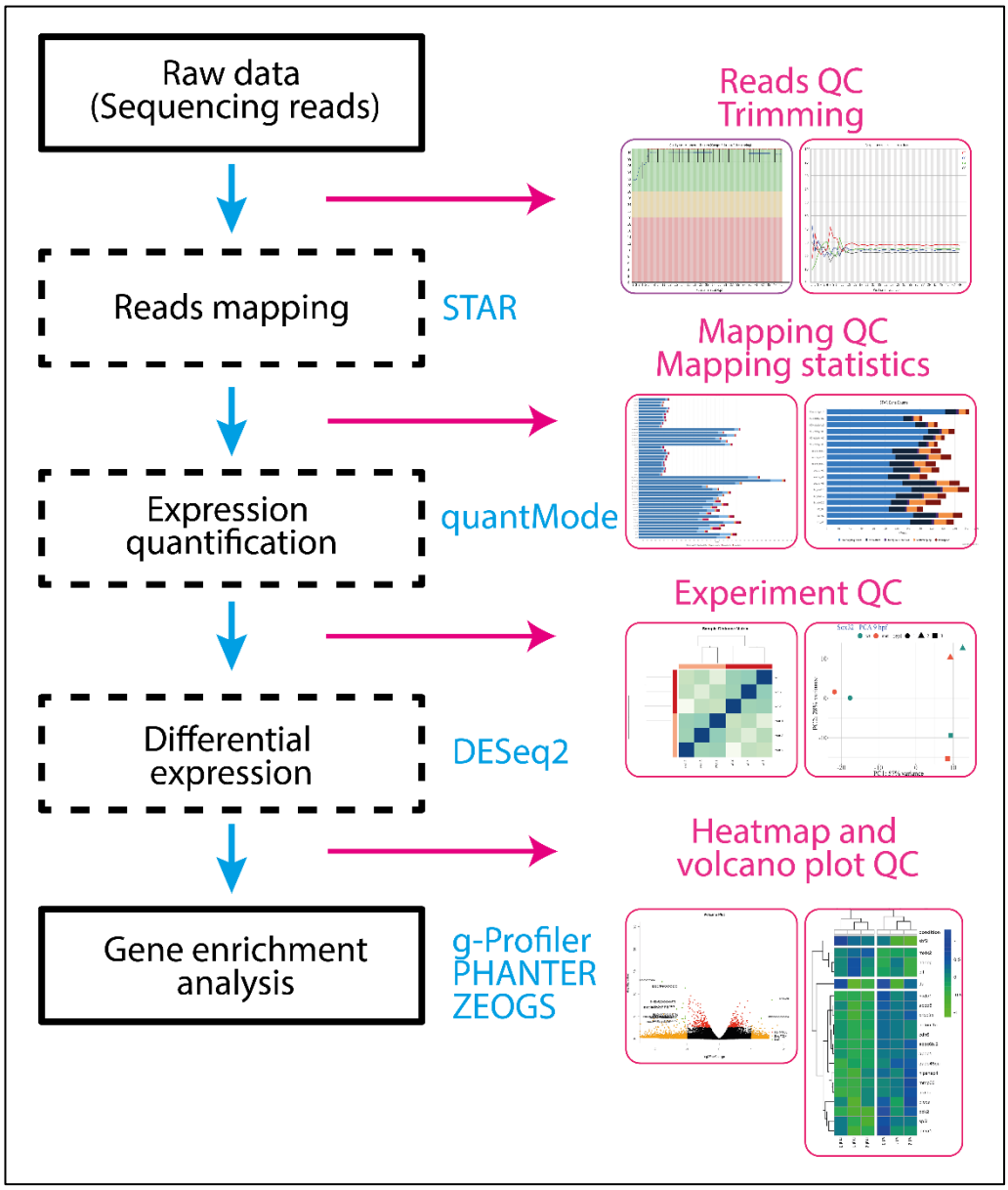


Figure 5.2 Bioinformatic workflow to analyse sequenced data (computational phase). Schematic representation of the workflow from obtaining the raw data to gene enrichment analysis. Solid boxes denote start and end points; dashed boxes intermediate steps. Blue text denotes methodologies used for the respective step; pink text denotes quality control steps (see Materials and Methods for full details).

5.3 Quality Control in RNA Sequencing - wet-lab phase

RNA-seq has increased our knowledge of the breadth and depth of eukaryotic transcriptomes, however, as with all NGS technologies artefacts can be introduced into these data. Such artefacts can compromise and confound the key biological meaning of data from high-throughput sequencing technologies. In particular, RNA-seq experiments suffer from intrinsic biases such as GC content and nucleotide composition bias and species-specific transcriptome complexity further affects data interpretation making it more imperfect (Altman

and Krzywinski, 2015; Schwartz et al., 2011). This is primarily because RNA-seq experiments are a sum of complicated, multistep processes involving purification, fragmentation, reverse transcription, ligation with adapters, PCR amplification and sequencing, and bias can be added in any step, culminating in unusable end-point data. The first part of this workflow describes the quality control metrics most often used, both from the point of view of library preparation (RNA quality, library enrichment) and bioinformatics approaches. The latter includes: (a) raw data sequence quality, sequencing depth, nucleotide composition bias and GC bias and (b) alignment quality, PCR bias and read duplication rates, contamination percentage (rRNA and mitochondria reads) and uniformity of coverage. A comprehensive quality assessment was the first step employed for all downstream analyses; an extended review of such quality assessment methodologies and software tools can be found in Wang et al. (2012), Consortium et al. (2014), Conesa et al. (2016), Sheng et al. (2017) and Zhou et al. (2018). The quality of the input material, the choice of library kit and the sequencing itself introduce bias which can compromise the downstream analysis. Variance in these techniques can result in key differences in transcriptomic profiles, leading to the misidentification of changes that have occurred due to experimental treatment rather than from biological differences. Different methods of input isolation and storage methodologies make direct comparison of RNA-seq data from different labs and protocols difficult and unreliable (Kukurba and Montgomery, 2015; Wang et al., 2018).

Taking the above into account, strict protocols and quality control (QC) are key to successfully performing RNA-seq experiments. I therefore applied rigorous QC steps to allow meaningful biological values to be extrapolated from my experiments: i) QC checks on the starting nucleic acids (RNA quality), ii) QC checks on library preparation (Library quality) and iii) QC checks on post-sequencing reads (Read quality).

5.3.1 RNA extraction

Library preparation is an important part of the RNA-seq workflow and the currently available kit-based methodologies for library preparation offer streamlined protocols and optimised yields. However, the initial extraction of the RNA, its quality and accurate quantitation still remains critical to ensure successful cDNA synthesis and library construction. Multiple methods have been described regarding how best to extract and separate RNA from DNA and proteins in cells. Different RNA extraction methods result in different RNA populations as some methods result in the loss of tRNAs, 5S rRNAs, snoRNAs, and/or

other RNAs < 250–300 bp and this ultimately influences the RNA-seq data obtained and the subsequent downstream analysis (Kałużna et al., 2016; Sultan et al., 2014)

I first searched through the literature and identified that both silica gel membranes or liquid-liquid extractions with phenol-chloroform (TRIzol) were common methods for RNA extraction from zebrafish embryos (de Jong et al., 2010; Hostelley et al., 2017; Peterson and Freeman, 2009). Modification of protocols depended on the stage of the embryos, the number of embryos and what species of RNA the investigator was interested in. In respect of the latter, different assays were shown to selectively enrich for some species of RNA, for example enrichment of large RNAs or specific loss of small RNAs (17-200 nt), leading to bias in downstream analysis. When using silica gel membrane extraction in column, selectively bound RNA molecules remain adhered to the column whilst the remaining cellular components are washed away through different cleaning steps. Ethanol is required in all kits using silica-gel membranes for isolating RNAs, and the first step of the protocol, where a volume ratio of ethanol is added to the sample, impacts the species of transcripts that will bind to the membrane, with higher ethanol volume ratios (1:1) resulting in the retention of RNAs < 200 bp. No phase separation, nucleic acid precipitation, or post-purification steps are then required. The second commonly used methodology, isolation of RNA by phase separation using phenol-chloroform (TRIzol) extraction, also effectively isolates RNA from a variety of sample sources. In this method, the cellular components are separated into 3 phases: the organic phase (proteins), the interphase (DNA) and the aqueous phase (RNA). Phenol-chloroform (TRIzol) extraction can be followed by either an alcohol precipitation (ethanol or isopropanol) to desalt and concentrate the RNA, or by an RNA concentrator column step.

For both extraction methods used, I performed additional purification steps by treating the RNA with DNase (in column). This step helped to remove any residual DNA as even small amounts of DNA contamination can negatively impact downstream results (NuGEN, 2013).

Both of the previously mentioned protocols resulted in good quality RNA yields and the samples were clear of proteins and organic contaminants which can inhibit the library making process (spectrophotometric analysis). However, I observed that when using significantly less starting material (sorted cells compared to whole embryos), the phenol-chloroform (TRIzol) extraction technique performed noticeably better in terms of reduced DNA contamination and degradation of RNA (Chapter 4). Therefore, I chose to use the phenol-chloroform (TRIzol) extraction technique to isolate RNA for all my experimental samples, to ensure consistency of

approach both within and between my experiments. The full details for the TRIzol extraction technique can be found in Materials and Methods.

5.3.2 Quality control of RNA preparation

It is crucial to start with high quality RNA; compared to DNA, RNA is much more prone to degradation and the quality of the extracted RNA has been shown to strongly impact the results of both microarray and RNA-seq experiments (Gallego Romero et al., 2014). Not only does the use of degraded RNA result in low yields but also can lead to a total failure of the RNA-seq libraries. RNA integrity was therefore assessed both by visual inspection of the ribosomal RNA bands via gel electrophoresis (a cheap and user prone bias method) and by using the RNA Integrity Number (RIN) estimated by the Agilent Bioanalyzer. Agilent has developed a software algorithm that calculates a RIN from a digital representation of the size distribution of RNA molecules, which is therefore an objective measure of RNA quality.

The integrity and size distribution of total RNA isolated was checked by electrophoresis on an agarose gel and stained with ethidium bromide (Figure 5.3). 2 clear bands, the 28S and 18S rRNA (approximately 4200 bp and 1900 bp respectively) were visible on a gel in a roughly 2:1 ratio indicating the RNA isolated was of good qualit (Figure 5.3A). As RNA degrades, this 2:1 ratio decreases and low molecular weight RNA becomes detectable, as shown in the first lane (Figure 5.3A). Completely degraded RNA would appear as a very low molecular weight smear (not shown).

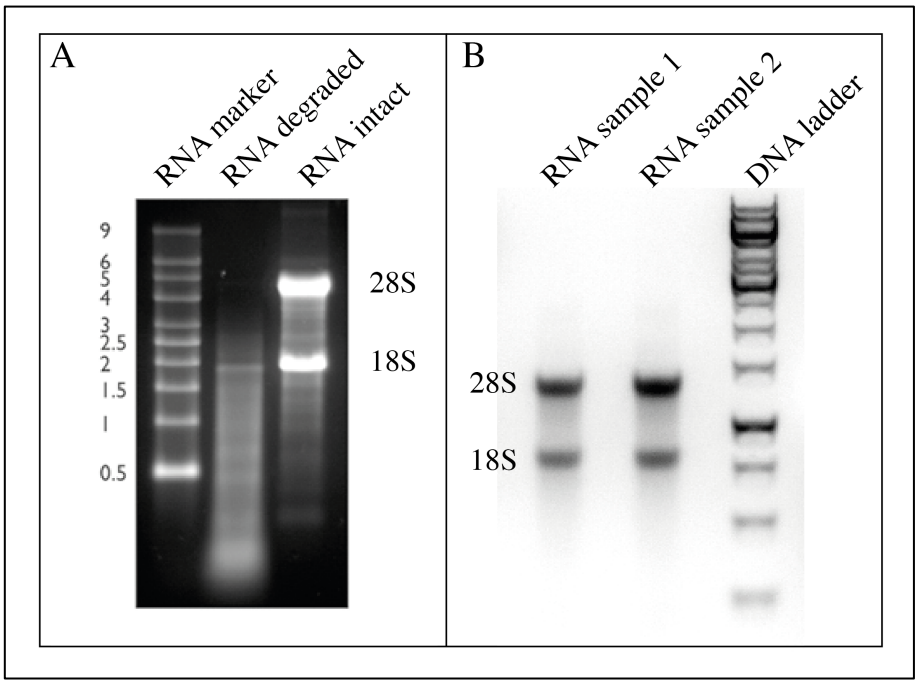


Figure 5.3 RNA gel electropherogram. (A) RNA integrity assessment was based on the ratio of 28S rRNA to 18S rRNA; good quality RNA should show a roughly 2:1 28S rRNA:18S rRNA intensity (lane 2). Degraded RNA showed a smeared appearance (lane 1). Image adapted from Thermo Fisher, 2018. **(B)** RNA samples 1 and 2 are representative of the RNA isolation I performed, confirming the RNA I extracted was of good quality. DNA ladder shown is for reference purposes only and does not indicate band size.

Since the interpretation of gel images is subjective and has been shown to be inconsistent, it is recommend using a second method to determine RNA quality, such as the RIN estimated by the Agilent Bioanalyzer (Garcia-Elias et al., 2017; Reiman et al., 2017; Wimmer et al., 2018). This capillary electrophoresis approach simplifies the interpretation and reproducibility of RNA quality evaluations by providing a RIN, a numeric scale from 1 to 10 where 1 is the most degraded and 10 is the highest quality sample. A RIN allows one to compare the quality of RNA from different samples, on different extraction days and using different kits in a standardized manner; RNA used for RNA-seq experiments should be as intact as possible with a minimum RIN of 7. The extraction protocol I used (see above) allowed me to obtain consistently high RIN scores (7-10) (Figure 5.4B-D). I reisolated RNA from samples with low RIN values (6 or below) or where large outliers from the average RIN of a group of samples was present. Only the RNA samples shipped from Austria had a RIN < 6 (Figure 5.4A). I observed that the RIN increased with older embryos; RNA extracted at 5.25 hpf was more degraded than at 9.00 hpf (Figure 5.4B,C) and also that using a higher number of embryos per sample gave higher average RIN values (data not shown).

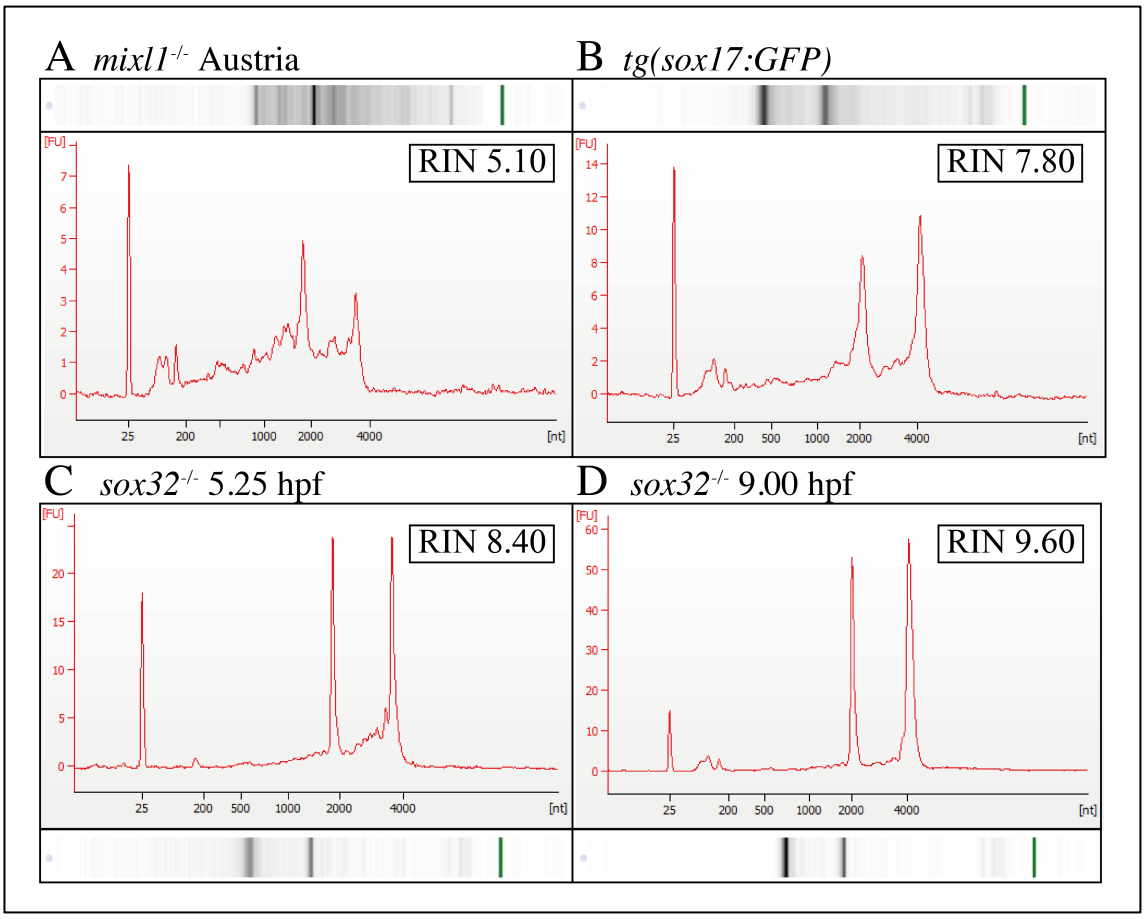


Figure 5.4 RNA capillary electropherogram. RNA integrity assessment was based on the ratio of 28S/18S rRNA. 1µl of RNA was run in the Bioanalyzer chip producing RIN. (**A**) RNA extracted from Austrian samples with showing RIN < 6 (degraded RNA), (**B**) good quality RNA from sorted cells . (**C**) and (**D**) showed good quality RNA from *sox32*$^{-/-}$ 5.25 hpf and 9.00 hpf embryos respectively. Note that RNA integrity was higher from older embryos (compare D to C).

I then checked that the extracted RNA was free from carryover genomic DNA (gDNA) contamination using PCR. I design intron spanning primers thus different band sizes were observable between gDNA (larger fragment due to the presence of an intron) and the respective cDNA control (smaller fragment containing only the coding sequences) (Figure 5.5). DNase treatment of the purified RNA with RNase-free DNase was recommended (NuGEN, 2013) hence, I tested different digestion incubation times and determined that in-column DNAse treatment for 10 mins at 37°C was optimal for eliminating DNA contamination whilst retaining RNA quality.

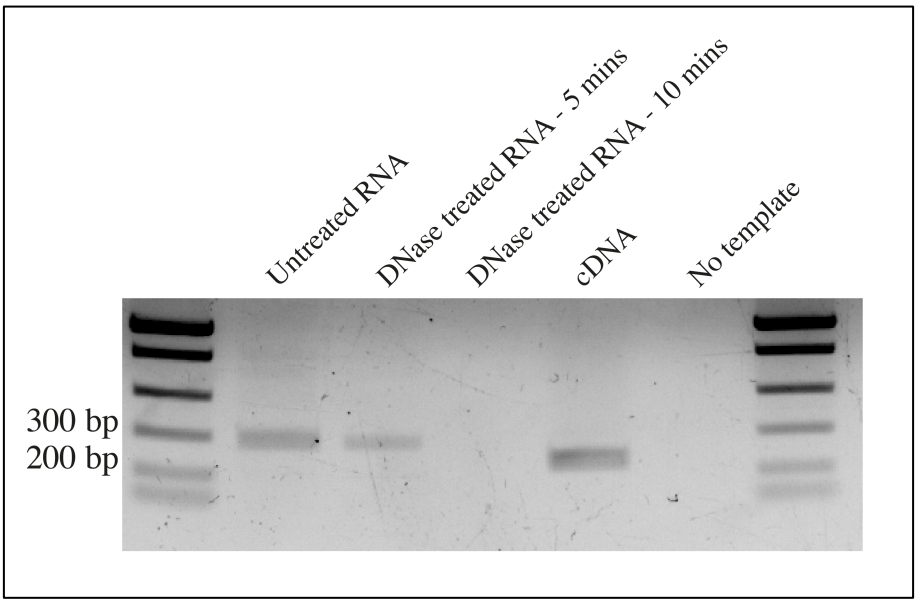


Figure 5.5 Example of the gDNA contamination PCR. Gel image illustrating gDNA contamination in RNA samples from sorted GFP$^+$ cells with and without in column DNase I treatment as described. PCR was undertaken using 18S intron spanning primers thus yielding different sized products from cDNA (smaller band) and gDNA (larger bands) respectively as shown. No DNA contamination was detectable after 10 min of DNase treatment.

5.3.3 Quantification of RNA

RNA quality is not the only important metric to evaluate when preparing an RNA-seq library, the concentration is also an important quality check, as different library kits from different companies are stringent on the amount of starting point input. It was therefore important to accurately quantify my RNA samples to ensure that the sample input was of sufficient yield to generate reproducible data. I used a robust and non-subjective strategy for RNA quantification comprised of spectrophotometry to assess contamination and carryover from the RNA extraction method, fluorometric method (Qubit with specific intercalating fluorophores) and gel based microfluidic electrogram to determine the quantity of RNA. An initial estimate of RNA concentration was obtained by measuring absorbance at 260 nm with a DenoVix spectrophotometer, and because sample dilution was not necessary, pipetting and dilution errors had no effect on correct determination of concentration. In addition, the DenoVix readout provides an estimate of contaminant carryover (e.g. phenols from TRIzol extraction). RNA purity was determined by evaluating the 260/280 and 260/230 ratios; high absorbance in the 280 nm range indicates the presence of proteins while excessive absorbance at 230 nm denotes the presence of residual carryover of phenols in the sample from the extraction step. The 260/280 ratio for my RNA samples was approximately 2.0 and the

260/230 ratios were between 1.8-2.2. However, this measurement was reinforced using the other quantitative methodologies, as it should be noted that all nucleic acids have a peak absorbance at approximately 250-260 nm, including RNA, DNA, and free nucleotides. When RNA preparations contain DNA or free nucleotides it affects the ability to accurately determine the RNA concentration using a spectrophotometer. Therefore, RNA concentration was also assessed using 2 other techniques: Bioanalyzer and Qubit. I observed that when the Nanodrop was properly calibrated, there was less than 15% variance in concentration measured by all 3 methods. For any samples where the Nanodrop ratios differed significantly from those detailed above, samples were either repurified or RNA reextracted until sufficient ratios were observed.

5.3.4 Quality control of RNA library

Due to the several types of RNA, a variety of library preparation enrichment protocols are available and, depending on the researcher's specific question, selection of the appropriate library preparation protocol should be guided by the study objective(s). In high-throughput sequencing terms, a library is defined as a random collection of DNA fragments with adapters at the end that are ready for sequencing; when preparing ChIP-seq libraries for example, gDNA bound by the protein of interest is fragmented and used to generate the library. For RNA-seq libraries, oligo-dT enriched RNA or ribosome depleted RNA fragments are converted to cDNA libraries. These cDNA libraries are comprised of fragments that are typically between 200 and 600 bp long (including the adaptor sequence ~ 65 bp), and after hybridization to the flowcell (via the adaptor sequence), the ends of the fragments are sequenced either from only one end (single end sequencing; SE) or both ends (pair end sequencing; PE). RNA-seq libraries are usually PE sequenced, the 2 most important reasons for this are as follows: 1) The ability to detect gene fusions and characterise novel splice isoforms. PE sequencing provides better resolution of the 3'-end of the transcript and thus is better at defining 3'-UTRs and novel ncRNAs. 2) PE sequencing estimates the size of the insert and aiding in the prediction of deletions, mutations and inversions within the genome (Conesa et al., 2016).

The next decision I faced was which sequencing protocol to use for my libraries; more specifically, which RNA enrichment method to employ and whether to use stranded or non-stranded RNA-seq transcriptome profiling. In respect of the RNA enrichment method, since the advent of RNA-seq technology a decade ago, comparisons between rRNA depletion

methods and polyA+ selection have been evaluated by various independent researchers using samples, cell lines, model organisms and different protocols/kits (Sultan et al., 2014; Wang et al., 2016; Zhao et al., 2018; Zhao et al., 2014b)

After extraction, total RNA contains ribosomal RNA which comprises the majority of it (> 80 to 90%) and these highly abundant rRNA molecules (which are of little interest in this context) must be removed from the samples before sequencing to allow for efficient detection of transcripts/gene proportions. The 2 routine approaches are either the depletion of rRNAs using affinity probes with complementary rRNA sequences or capture of polyadenylated RNA (polyA+) transcripts using oligo-dT primers. Both methods have distinct advantages and limitations. Studies that require low sequencing depth and where the main focus is on the protein-coding fraction of a transcriptome usually opt for polyA+ selection. In contrast, rRNA depletion allows the efficient removal of both cytoplasmic (45S, 28S, 18S, 5S) and mitochondrial rRNA transcripts in human, mouse and rat samples, thus preserving all other relevant RNA species. rRNA depletion has therefore been used for most transcriptomic studies that take a more comprehensive view of transcriptome composition (lncRNA, miRNA etc.) (Sultan et al., 2014; Zhao et al., 2018; Zhao et al., 2014b). polyA+ selection is the faster method, with less hands-on steps, but it provides information only for mature transcripts and the samples need to be of high quality, with intact poly-A tails. Degradation of poly-a tails introduces bias in the selected transcripts and can skew downstream analysis (Kukurba and Montgomery, 2015). Another important consideration was that rRNA-depletion library kits cost significantly more than polyA+ selection libraries, and the resulting libraries obtained need more sequencing depth to have a comparable coverage of protein-coding reads (compared to a transcriptome from polyA+ enriched transcripts). Finally, a key technical advantage that favours rRNA depleted libraries compared with polyA-selected libraries is that the former perform better for degraded RNAs, as they do not rely on an intact polyA+ tail (in particular, polyA+ would have not been applicable to Austria *mixl1*$^{-/-}$ samples).

Once all isolated RNA samples had passed the aforementioned quality control measures, I performed either rRNA depletion or mRNA enrichment. Specifically, all RNA samples with a RIN value higher than 8 were subject to polyA+ (mRNA) enrichment, whereas RNA samples with a RIN value between 4.5 and 6 were subject to rRNA depletion using the RiboZero enrichment protocol from Illumina. As the Austrian *mixl1*$^{-/-}$ samples had RIN values of < 6 and were therefore subjected to RiboZero enrichment, I also chose to use the same protocol

for my *mixl1*$^{-/-}$ samples, as despite them having RIN values of ≥ 7, I wanted to reduce variability between the *mixl1*$^{-/-}$ samples as much as possible.

Following depletion/enrichment, the RiboZero enrichment samples were again assessed on the Bioanalyzer to confirm the success of the depletion/enrichment by absence of the ribosomal RNA peaks (data not shown).

Early developed methods for RNA library preparation did not retain information on the DNA strand from which the RNA molecules were transcribed. Only more recently has it become possible to retain the strand information by modifying the standard RNA-seq protocol, known as strand specific or stranded RNA-seq (Zhao et al., 2015). The ability to obtain information on the originating strand is useful for many reasons, including the identification of antisense transcripts, determination of the transcribed strand of non-coding RNAs and determination of expression levels of coding or non-coding overlapping transcripts. Overall, the ability to determine the originating strand can substantially enhance the value of RNA-seq experiments.

The final feature linked to library quality that I needed to consider before shipping my libraries for sequencing was the size insert of the generated libraries (i.e. the fragment size). Purified libraries were analysed using the 2100 Bioanalyzer with a High Sensitivity DNA Chip and the insert size was verified as being within the expected range of 300-700 bp, with no contamination of adapter-dimers or PCR duplicates (Figure 5.6 B). The latter can be a major issue on the new clustering chemistry for Illumina HiSeq4000 (as used for my samples) and the most recent HiSeqX Ten (Illumina FAQs). Adaptor dimers present a problem because they are sequenced much more readily than the longer library fragments, causing a significant reduction in relevant reads. Adapter dimers can be minimized by optimizing the adapter:insert ratio during library construction or by re-purification using AMPure XP beads procedure (NEBNext Ultra II RNA Library manual). I adopted the latter for the few libraries (primarily *mixl1*$^{-/-}$), that showed adaptor dimer contamination (Figure 5.6A).

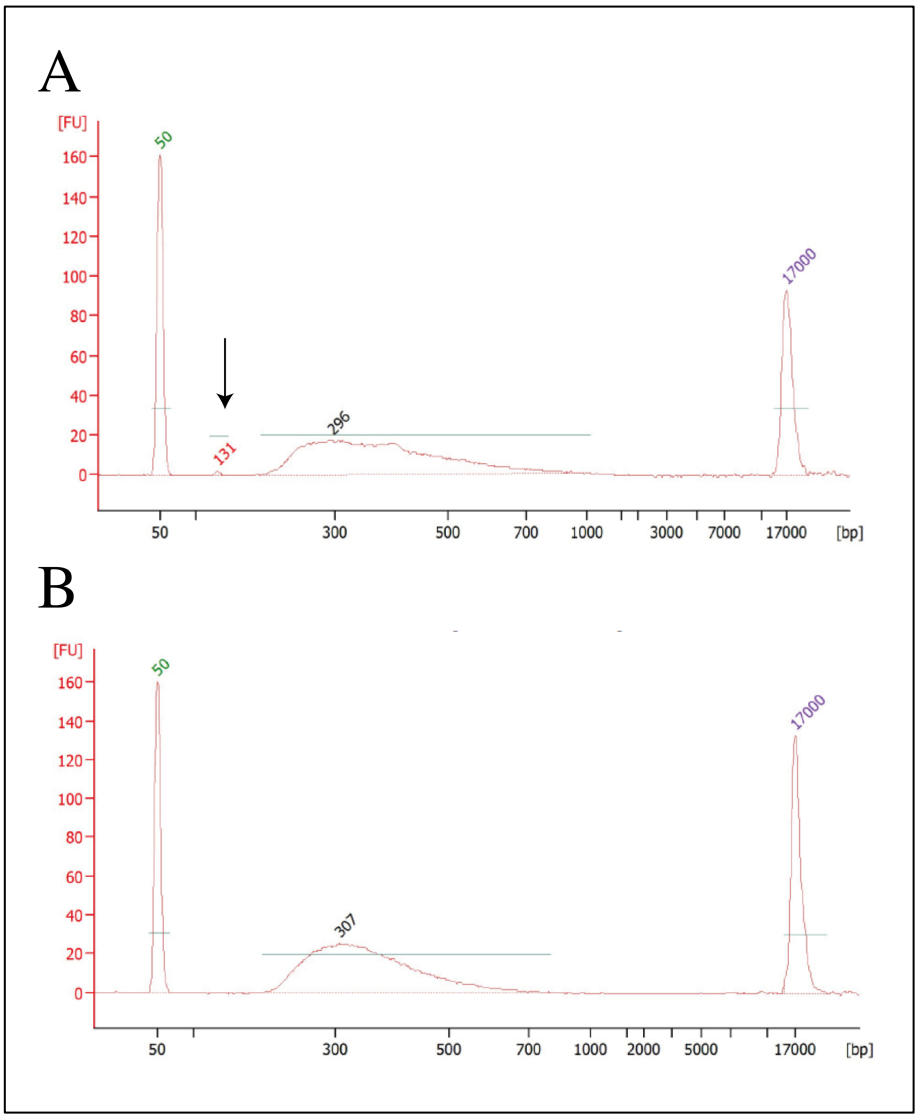


Figure 5.6 Representative Bioanalyzer profiles of RNA-seq libraries. (**A**) Bioanalyzer trace of a *mixl1*$^{-/-}$ library with adapter dimers contamination at around 120 bp (black arrow). This library was repurified as described in the text to remove the contamination. (**B**) Bioanalyzer trace of a reselected, purified library submitted for sequencing showing a single peak at the expected molecular weight (around 300 bases). Lower and upper markers at 50 and 17,000 bp are shown. FU: fluorescence unit.

In summary, my key aim was to identify and quantify any differences in expression of mRNA associated with the genotypes of the fish in my study (*sox32*$^{-/-}$ and *mixl1*$^{-/-}$), compared to wild-type. More specifically, I wanted to determine what changes the mutations caused in the expression of protein-coding genes, therefore where my RNA samples were of high quality, I enriched for mRNA using the polyA+ enrichment method. The *mixl1*$^{-/-}$ RNA extracted from the samples shipped from Austria had low RIN values (< 6), therefore in order to use these samples I decided to prepare rRNA depleted libraries. The RNA that I extracted from the *mixl1*$^{-/-}$ line in London was also done using the rRNA depletion protocol in order to

combine the repeated measurements on *mixl1*$^{-/-}$ embryos. Finally, I decided to preserve the strand information as stranded RNA-seq provides a more accurate estimate of transcript expression levels compared with non-stranded RNA-seq, and multiple sources recommend this approach for mRNA-seq studies (Kukurba and Montgomery, 2015; Zhao et al., 2015).

5.4 Quality Control in RNA Sequencing - computational phase

The aim of the bioinformatic analysis was to find genes that were differentially expressed between WT and mutant embryos and to discover new markers in the enriched endodermal GFP^{+} population. The first steps depended on a command-line interface using shell scripting language and were the most computationally demanding, therefore were performed on Rosalind, a King's College High Performance Compute Cluster (HPC) (Figure 5.2 – Reads mapping).

The second part of the analyses was carried out using R programming language and RStudio to perform statistics and visualisation. In particular the statistical analysis process included data normalisation, graphical exploration of raw and normalised data, test for differential expression for each feature between the conditions, raw p-value adjustment and export of features having a significantly differential expression between the conditions (Figure 5.2 – expression quantification and differential expression).

Quality control and data quality assessment were essential steps of the data analysis, I addressed each data set individually and removed data with insufficient quality early in the analysis, and then proceeded to differential expression testing. I first outline the main pipeline and later on present the individual results accounting for variation of the streamlined protocol. I define the term quality as fitness for purpose, meaning that quality was the pragmatic interpretation/detection of differentially expressed genes, and I exclude samples whose experimental treatment suffered from an anomality that reduced the statistical power or confounded the results/data points.

5.4.1 Data Records and quality control

Thirty-six raw FASTQ sequencing files were retrieved from BGI sequencing facility, the 49 bp reads were generated from 2 lanes on Illumina HiSeq4000. I performed an initial QC checks using the FastQC software, a general NGS QC package that is applied before primary

biological analysis. This step is similar to the analysis I did with the raw files from the ChIP-exo experiments described in Chapter 3. The purpose of this initial QC assessment was to inform me about the quality of the sequencing chemistry, whether the sequence reads required 'trimming' to remove low quality bases at the beginning or at end of the read, and whether the data required trimming to remove sequencing adapters. Incorrectly called bases and adaptors negatively impact assembles, mapping, and downstream bioinformatics analyses and therefore have to be removed before further analysis. Out of the 11 plots available in the FastQC report and the statistics, the following were the most informational: per base sequence quality, per sequence quality scores, per base sequence content, sequence duplication levels and overrepresented sequences adapter content (Figure 5.7). The per base sequence quality plot (Figure 5.7A) represents the quality of a base pair linked to its position in the read; modern sequencing technologies produce reads that have deteriorating quality either towards the 3'-end or towards the 5'-end and some in both. This is related to the chemistry used by Illumina sequencers; higher average quality is spotted in bases in the earliest rather than in later cycles of the sequencing procedure. As you can observe in the QC plot below, the quality of a base pair did not decrease with the length of the read (reads were only 49 bp long). The per base GC content plot in panel B of Figure 5.7 represents the average proportion of single bases (A, C, G and T) along the length of the read (1 to 49 bp). A bias was observed in the first 11 bases which is attributed to RNA-seq protocol steps, as discussed below. This contrasts with other sequencing technique such as Chip-seq/ChIP-exo data or ATAC-seq (Figure 3.12A) which do not show this positional bias of the reads. This bias had most likely been introduced during the conversion of RNA to cDNA. Random hexamer primers during reverse transcription, as used by most library kits such as NEBNext, Illumina TruSeq and Kapa biosystems, together with the specificity of the polymerase to start transcribing from particular regions, and artifacts from end repair, could all account for introduction of the observed bias. For non-stranded specific RNA-Seq data, the average amount of all 4 nucleotides should be similar at any position within reads, whereas an enrichment for T over A nucleotide can be observed in dataset where RNA was selected using poly-dT beads. Panel C showed per sequence GC content, the GC content is shown as a function of the position in the read, where proportion of G+C of the sample over the curve should match the theoretical distribution (blue curve) of GC content, which in my libraries matched closely. The last important plot to evaluate was sequence duplication level as shown in panel D, Figure 5.7. This plot denoted the level of duplicate sequences in the library, which plots the proportion of total reads over the read sharing the same starting and ending base. It was used as an estimate of how many PCR

duplicates were present in the library, e.g. too many PCR cycle during the last amplification step in library preparation. As a result, it is common to observe high duplication levels for sequences originating from highly expressed genes. Duplication percentage in my libraries varied from 16 to 48%. Although I followed the recommended numbers of PCR cycles for the library amplification step, the suggested number is < 25%.

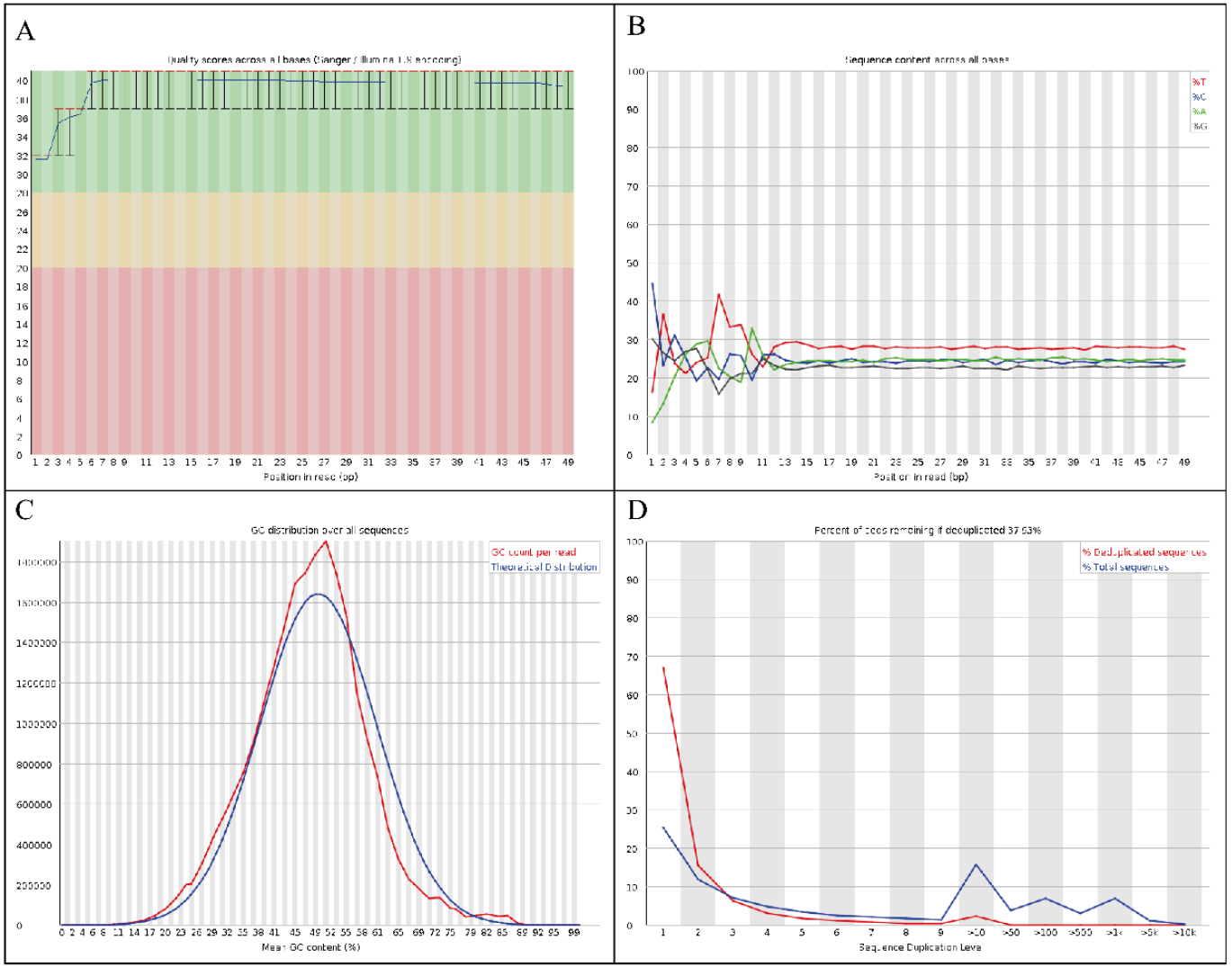


Figure 5.7 Quality checks from FastQC report of an RNA sequencing sample. (**A**) shows the quality of the bases as a function of the position in the read. (**B**) shows the average proportion of individual bases (A, C, G and T) spread across the length of the reads. The 11 bases at the 5' end of the reads showed a biased distribution. **(C**)

showed GC content over all sequences. The proportion of G+C should match the expected GC content of the sample. (**D**) Percent of duplicated sequences, overrepresented sequences which showed any over abundant sequence that was present in all the reads.

5.4.2 Filtered and trimmed data: quality trimming and adapter removal

Once the quality of the raw files was inspected, next step was to take the raw reads contained in the FASTQ file and trim them to remove low quality bases and the adapters before the mapping step; the step was performed to check whether mapping percent could be increased. I used TrimGalore, a wrapper around Cutadapt program, to remove the Illumina Sequencing adapter (AGATCGGAAGAGC) at the 3' end (Krueger, 2012). Additionally, 11 bases were also trimmed from the 5' end of the reads of all samples as they showed a biased distribution. The parameters used were similar to the ones described in Chapter 3 during quality assessment of ChIP-exo libraries with the addition of the PE parameter. For clipping the adaptor and trimming the reads, the parameters were set such that reads were trimmed when the average quality over a 5 bp window drops below 20, starting from the 5' end side of the read. The stringency of these parameters was confirmed by a posteriori QC validation (see below). Moreover, only reads that were at least 25 bp long after the trimming and quality removal were kept. This were simply because shorter sequences are harder to align, are more likely to be multi-mapped and are more prone to have originated from technical artefacts.

I then assessed the filtered and trimmed data again with a QC assessment by FastQC to ensure that the previous quality trimming and/or adapter removal steps effectively conserved high quality reads without being too stringent and without introducing any new arbitrary technical biases, which would be more detrimental than simply aligning the unprocessed read. As shown in Figure 5.8 below, changes were observed after the trimming step, in particular sequencing adapters were no longer identified as over-represented (the adapter removal effect), the distribution of sequence length shifted from all reads being 49 bp to reads with a range from 25 to 49 bp. As there is no gold standard method for processing sequencing data, I performed the process on all datasets, in particular as the raw reads from the Austria *mixl1*$^{-/-}$ libraries retained overrepresented sequences even after the trimming step, possibly indicating that an additional contamination was still present. After further investigating the datasets and adjusting the parameters of TrimGalore, Austria *mixl1*$^{-/-}$ libraries sequencing adapters were no longer identified and the data was of sufficient quality for mapping.

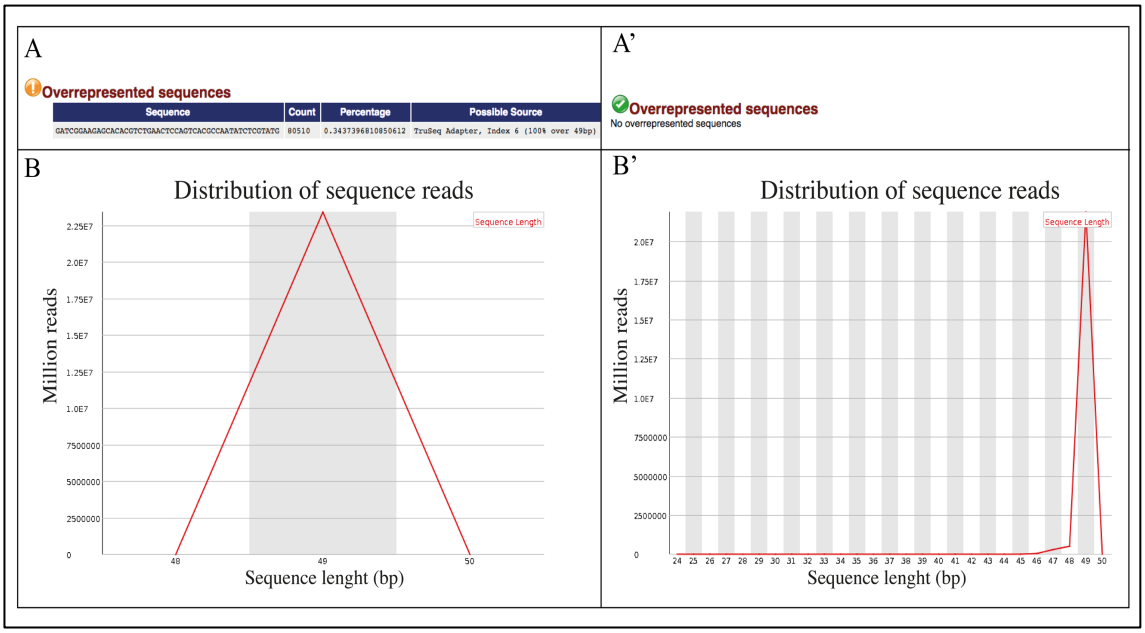


Figure 5.8 Quality check before and after trimming. QC plots extracted from FastQC reports at different stages of the data preprocessing. (**A)** Overrepresented sequences of Illumina adapter 6 was recognised by FastQC in 80510 reads. (**A'**) The same data after the trimming step. The adapter contamination was solved. (**B**) Sequence length distribution prior to trimming where only one peak at 49 bp was visible. (**B'**) After trimming the distribution of read lengths spanned from 25 to 49 bp.

These summarise and conclude the "technical" QC which inspected the raw data for technical biases due to sequencing (adapter contamination, base call quality issues, etc.).

5.4.3 Aligning reads to a reference genome

Once the data was of sufficient quality, the next step of the pipeline was to align the reads against the zebrafish reference (genome or assembled transcriptome). There are numbers of software to perform read alignment and the choice of the most appropriate aligner depends on the system and analysis goals; benchmarking papers that detail the advantages and disadvantages of each alternative software can be found in the following reviews (Baruzzo et al., 2017; Conesa et al., 2016; Costa-Silva et al., 2017; Medina et al., 2016). Features like type and state of the reference genome, type of sequencing, read length, algorithm speed, accuracy and sensitivity in aligning as well computational hardware requirements are discussed. At the time of writing and analysing the data, the zebrafish community is mostly split, ranging from the TopHat2 to STAR (Dobin et al., 2013; Kim et al., 2013; Langmead and Salzberg, 2012); both are splice-aware software but have different mapping accuracy, speed and hardware requirements. Recently, studies have started using alternative pipelines, switching from a

genome-based alignment to use transcript abundance quantification methods such as Salmon to estimate abundances without aligning reads (Patro et al., 2017).

I used STAR, a widely used ultrafast read aligner to align the reads for all my experiment to the Ensembl release 93 zebrafish reference genome Zv10. This choice was supported by the results of (Dobin and Gingeras, 2016; Sahraeian et al., 2017) which showed that STAR address many of the challenges of RNA-seq data mapping. In addition to the default parameters, I took into consideration the minimum and maximum intron size of the zebrafish genome (--alignIntronMin=30 and --alignIntronMax=1050000), so that STAR does not try to align split reads across a distance greater than the longest intron. Lastly, I specified --quantMode GeneCounts option in order to count reads per gene while mapping. For the visualisation, normalised coverage BigWig tracks for RNA-seq data were generated from the resulting sorted and indexed BAM files (sorted sequentially per chromosome position) using bamCoverage from the deepTools package (Ramirez et al., 2014) with a window size of 500 bp, and normalisation by TPM.

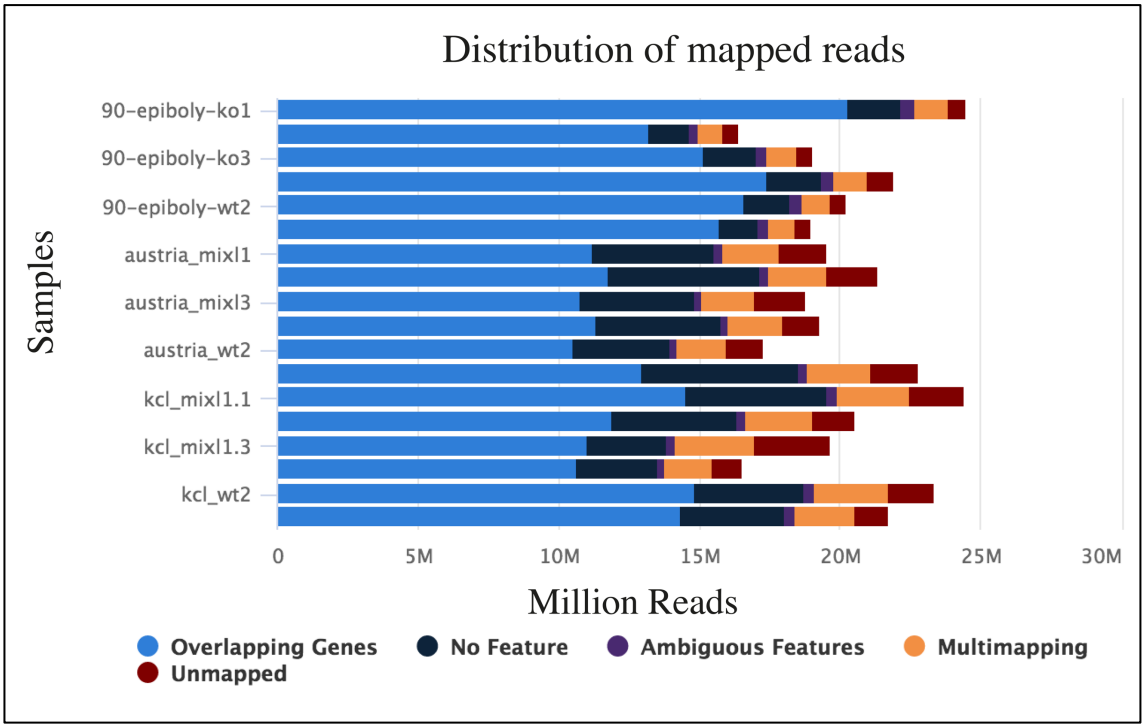


Figure 5.9 Graphical summary of mapped reads as output by MultiQC. Summary statistics of the mapped reads with STAR. Blue represents unique reads which were mapped to genes and were used for differential gene expression. Orange represents multi-mapped reads and red stands for unmapped reads. Note kcl_wt1 sample in which I inverted the strand information during the mapping process. The use of complementary plots throughout the bioinformatic pipeline as quality checks helped me interpreting

inconsistencies between samples and solved promptly possible problems. The sample was re-map correctly before downstream analysis.

Previously studies in human and mouse tissues has shown that 3 million reads were enough to identify differential genes (Ramsköld et al., 2012). In addition, Mortazavi et al. (2008) in their recommendations when designing RNA-seq experiments, concluded that cDNA sequencing with 10–40 millions are appropriate to identify the majority of isoforms expressed in mouse. All my samples have at least 10M unique mapped reads.

Read alignment concluded the data pre-processing steps common to the majority of RNA-seq experiments.

5.4.4 Analysis in R - Data pre-processing

Having obtained the mapping of the RNA-Seq reads to the genome, the consequent analysis steps to be performed will be determined by the project goals and the scientific questions that one wishes to address, probably the most common downstream analysis options are to identify differential expression between conditions or sequence variants. Distinctly different analysis methods are required depending on whether interest lies in identifying sequence variants or in exploring expression level differences between samples groups i.e. differential expression (DE). Once the reads have been aligned, there are a number of tools that can be used to count the number of reads/fragments that can be assigned to genomic features for each sample, the 2 most utilized methods are the featureCounts (Liao et al., 2014) or htseq-count (Anders et al., 2015). An advantage of using STAR, is that that by using the -countmode parameter the software automatically generates count matrices, as described in the following section, counting fragments and ignoring multi-mapping reads restricted to the sense strand. The resulting matrix of read counts was analysed using R., producing smaller files which store estimated abundances, counts, and effective lengths per transcript. In order to produce correct counts, the correct column from the STAR output needs to be selected, column 1/2/3 in particular, depending on which type of library protocol was used. All our RNA-seq experiments were strand-specific, hence the third column from STAR output file was used as input in R and DESeq2 to compute the differential gene expression analysis (Love et al., 2014).

After loading the data, I first had a look at the raw data table itself. The data table contains one row per annotated feature and one column per sequenced sample. Row names of this table are feature IDs (unique identifiers). The table contains raw count values representing the

number of reads that map onto the features. For this project, there are 35117 features in the count data table (numbers of gene in zebrafish genome).

Table 5.1 Partial view of the count files. Total number of counted reads for each sample and each gene are reported.

	wt_1	wt_2	wt_3	mut_1	mut_2	mut_3
ENSDARG00000000001	47	30	53	42	40	38
ENSDARG00000000002	0	4	4	0	3	6
ENSDARG00000000018	1875	1891	1790	2007	2511	2493
ENSDARG00000000019	185	151	176	223	212	219
ENSDARG00000000068	44	64	85	28	30	31
ENSDARG00000000069	637	550	676	652	626	626

On average, 15 million reads were generated for each sample, which were sufficient to detect differentially expressed genes (Figure 5.10).

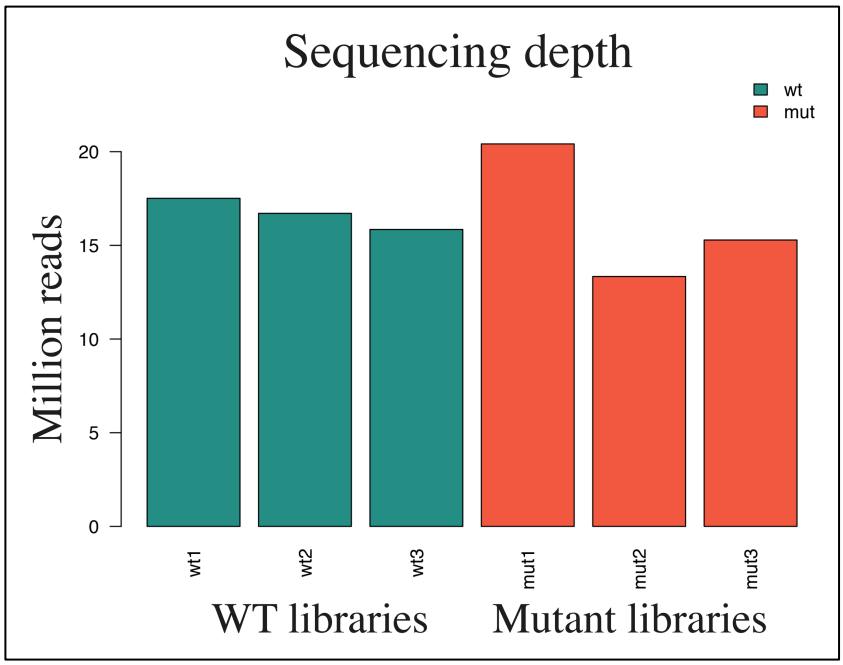


Figure 5.10 Example number of mapped reads. Total number of reads for each sample of the experiment. Colours match to the biological condition, WT in green and mutant in red.

I then inspect the percentage of genes with null read count in each sample, this percentage should be similar within conditions, meaning that the genes were not expressed/activated at that moment (Figure 5.11). In addition, genes with less than 5 counts were not taken into account for the analysis with DESeq2.

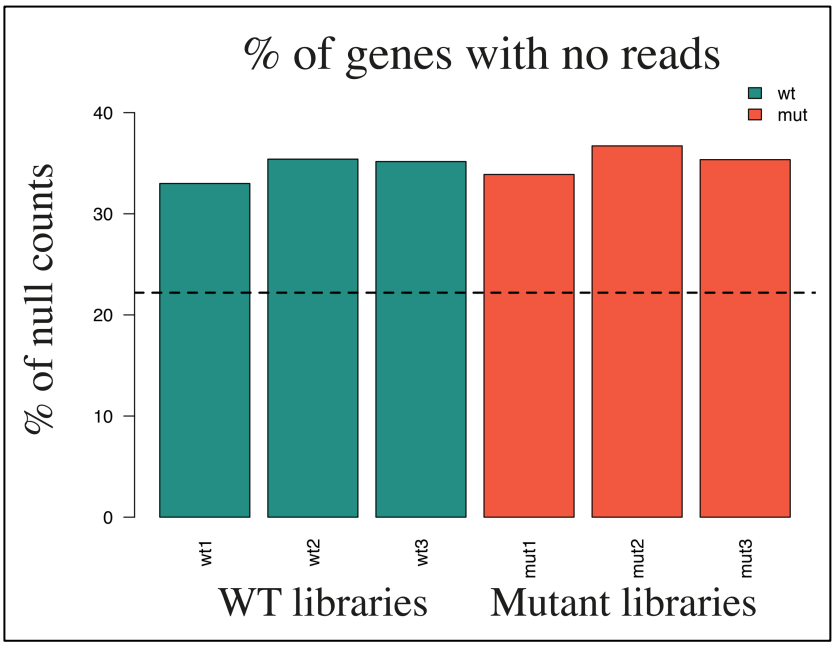


Figure 5.11 Example of percentage of genes with zero read counts in each sample. Genes which are not expressed in the control samples should match to mutant samples. Colours match to the biological condition, WT in green and mutant in red.

I next inspected the distribution of read counts for each sample. Raw counts were transformed $\log_2$(counts+1) to plot, and as before, replicates should have similar distributions (Figure 5.12).

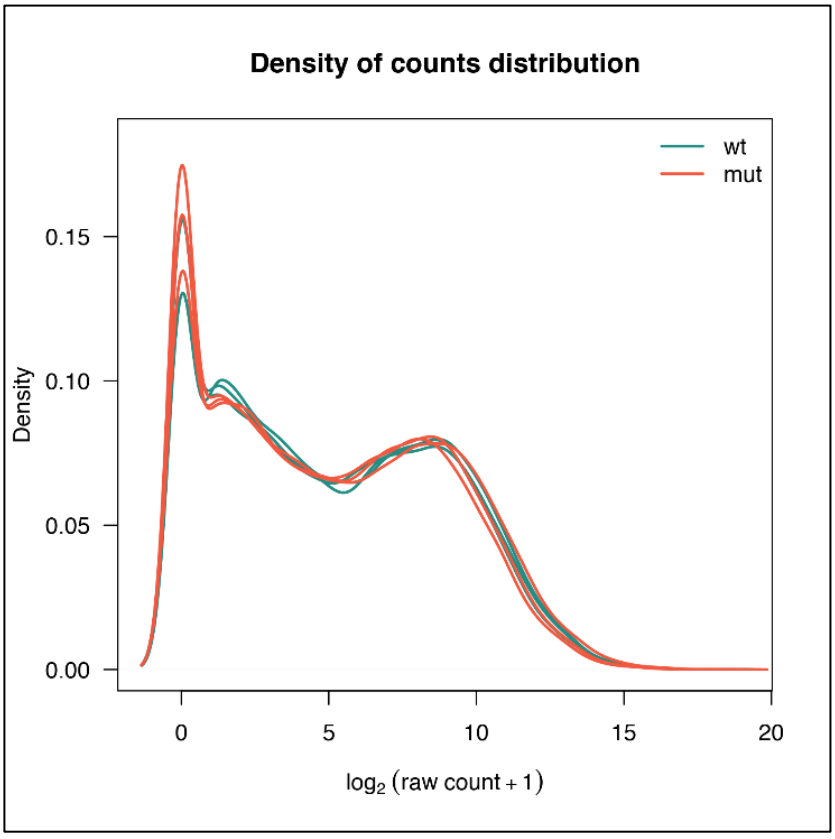


Figure 5.12: Density distribution of read counts. Average number of read per gene was around 8.

As described previously 2 approaches to selected and extract mRNA from a total RNA sample are either to deplete the sample of rRNA or selectively enriching the sample for poly-adenylated transcripts (rRNA is not poly-adenylated) with poly-dT beads, or to include a precipitation step that selectively precipitates only long (usually >200 bp) nucleotide fragments.

No protocol is sensitive enough to completely remove all rRNA and as a result, carryover of some rRNA is expected. This is not a problem *per se* as long as the percentage of the reads lost on rRNA is low, commonly between 0.1 and 5%. The contamination does not affect the downstream analysis if the libraries are sequenced at enough depth.

As a precautionary step, rRNA reads can be either filtered out before mapping with software such as SortMeRna originally developed to identify rRNA in metagenomics analyses (Kopylova et al., 2012) or the percentage of ribosomal reads can be evaluate after mapping. In the first approach, mapping rRNAs still produces valid alignment metrics and this can affect the overall alignment rate of the libraries, values < 70% mapped rate are considerate bad. Both methods inherently relied on the list of reference libraries containing the most common species of rRNA (5,16, 18, 23 and 28S) that I provided, this set is species-specific. In my libraries, the transcripts with most read came from housekeeping gene (ENSDARG00000020850) and mitochondrial gene (ENSDARG00000080337) which accounted for 2-4% in each library (Figure 5.13 and Table 5.2).

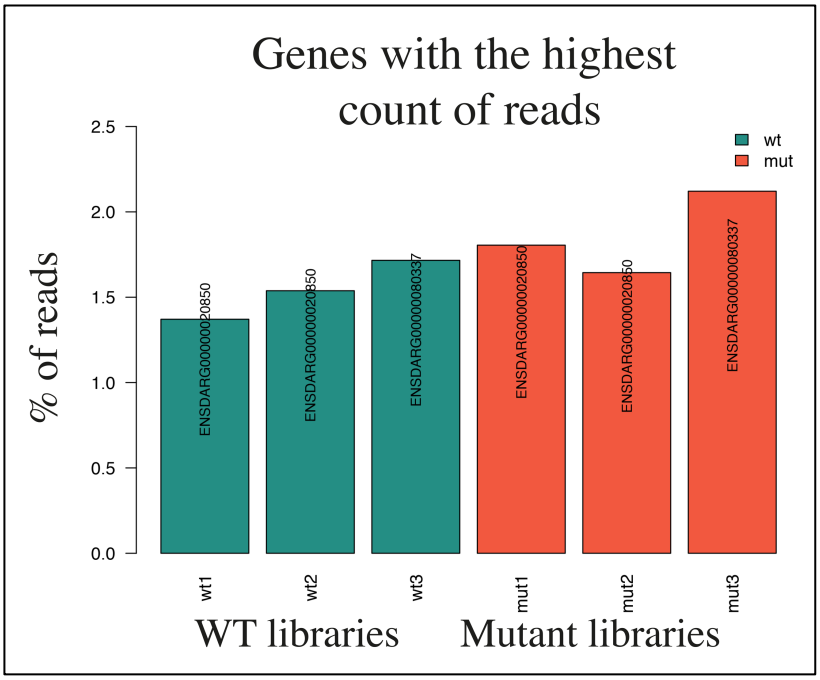


Figure 5.13 rRNA filtering. Percentage of reads associated with the gene having the highest count (provided in

each box on the graph) for each sample. ENSDARG00000080337 was a mitochondrial gene that fill between 2 – 4% of the reads.

Table 5.2: Genes with the highest percentage of read counts. A housekeeping gene and a mitochondrial gene were among the 3 genes with the higher count of reads.

		wt1	wt2	wt3	mut1	mut2	mut3
ENSDARG00000020850	*eef1a1l1*	1.37	1.54	1.59	1.8	1.64	1.2
ENSDARG00000028335	*hmga1a*	1.08	1.17	1.29	1.3	1.3	1.19
ENSDARG00000080337	*mt-nd3*	1.04	0.98	1.72	1.42	1.26	2.12

Subsequently, I assessed overall similarity between samples. I produced a pairwise scatter plot to show how replicates and samples from different biological conditions are similar or different. As for the density plot, $\log_2$(counts+1) was used instead of raw count values (Figure 5.14).

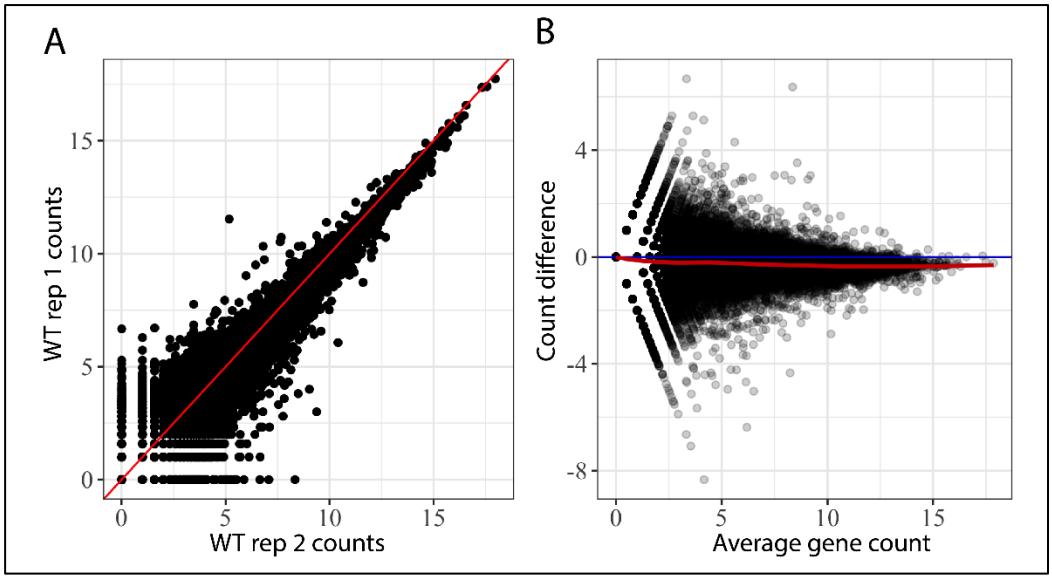


Figure 5.14 Reproducibility of replicates using the Pearson coefficient. (**A**) The data points are all concentrated near the line with small deviation, suggesting that the results were highly reproducible. (B) Counts difference between the 2 sample. WT 1 was used as reference (blue lines) whereas red line shows the difference in count numbers in the WT2 sample.

The same method was applied to all samples of an experiment to examine the reproducibility of biological replicates as shown in Figure 5.15 using the SERE coefficients (Schulze et al., 2012).

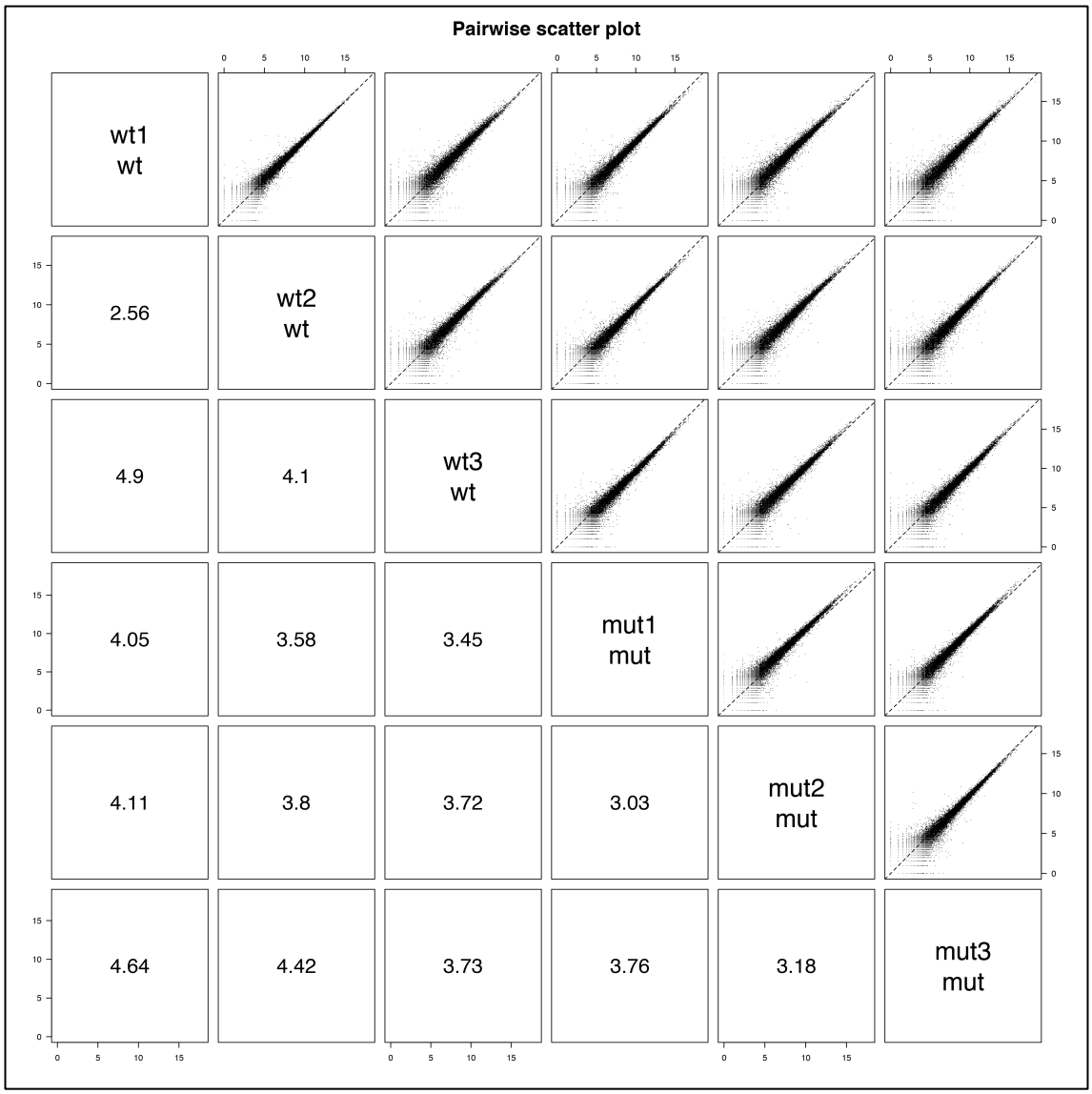


Figure 5.15 Pairwise comparison of samples using SERE coefficients. SERE coefficients are reported for each pairwise comparison. Higher similarity between the samples corresponded to lower coefficient values.

RNA-Seq data has a very large dynamic range of transcripts, reads from highly transcribed genes tend to receive more reads and therefore are over-represented and similarly lowly expressed genes have low read coverage and tend to be under-represented. Therefore, before examining patterns of expression across multiple samples, expression values across samples need to be normalised and accounted for differences in RNA composition. DESeq2 uses "relative log expression" normalisation to correct systematic technical biases in the data, in order to make read counts comparable across samples. This is done calculating a scaling factor for each sample. Figure 5.16 shows that the scaling factors of DESeq2 for the libraries.

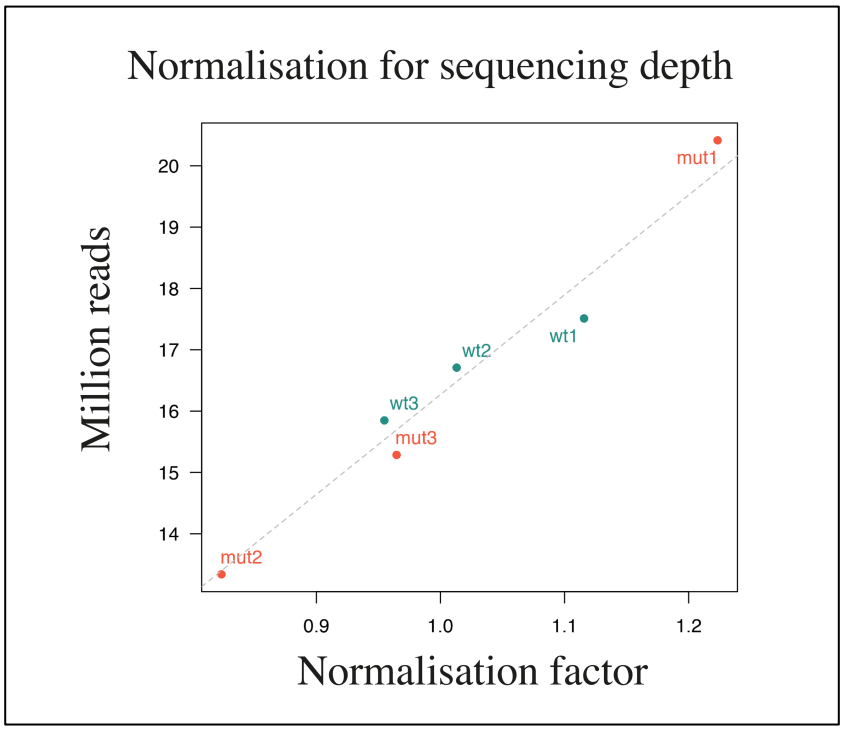


Figure 5.16 DESeq2 normalisation size factors. Read counts for each sample of the experiment are divided by a scaling factor. This normalises for sequencing depth and makes it more robust to compare the proportion of reads that were mapped to a gene in each sample. Plot of the estimated size factors and the total number of reads per sample; normalised counts are higher than row counts when the scaling factor is < 1, normalised counts are lower than raw ones when the scaling factor is > 1.

I then used boxplots to visualise the effect of the normalisation process, as they showed how distributions were globally affected during this process. Reads after normalisation should have a more similar mean across samples. Figure 5.17 shows boxplots of raw (left) and normalised (right) data respectively.

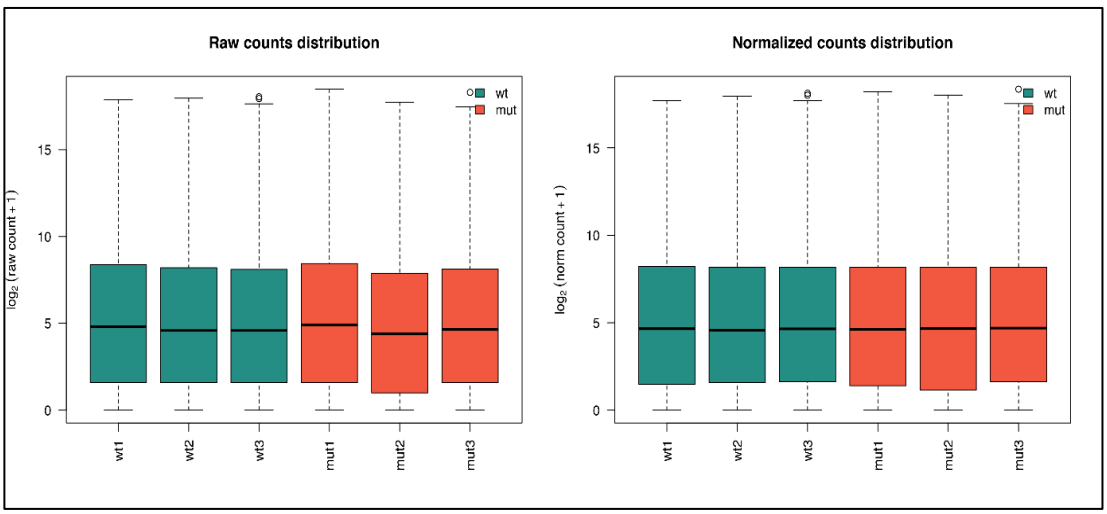


Figure 5.17 Boxplots of raw (left) and normalised (right) read counts.

Next step was exploring the datasets for "biological QC" meaning whether the observed effects/pattern in the data was related to biological causes and no obvious confounding factors such embryos collection date or sequencing lane were present. Through clustering the replicates and a PCA approach, I judged whether replicates were matching together by group and whether the condition (WT vs mutant and GFP^- vs GFP^+) looked sufficiently separated and then looked at the relationship between the samples.

I first performed a hierarchical clustering analysis of the replicates, where all transcripts were clustered by expression profile and sample-to-sample distances were calculated using the Pearson's correlation coefficient. Figure 5.18 displays the dendrogram obtained from the *sox32*$^{-/-}$ dataset, revealing good correlation of expression data within groups and low correlation between mutant and control WT groups.

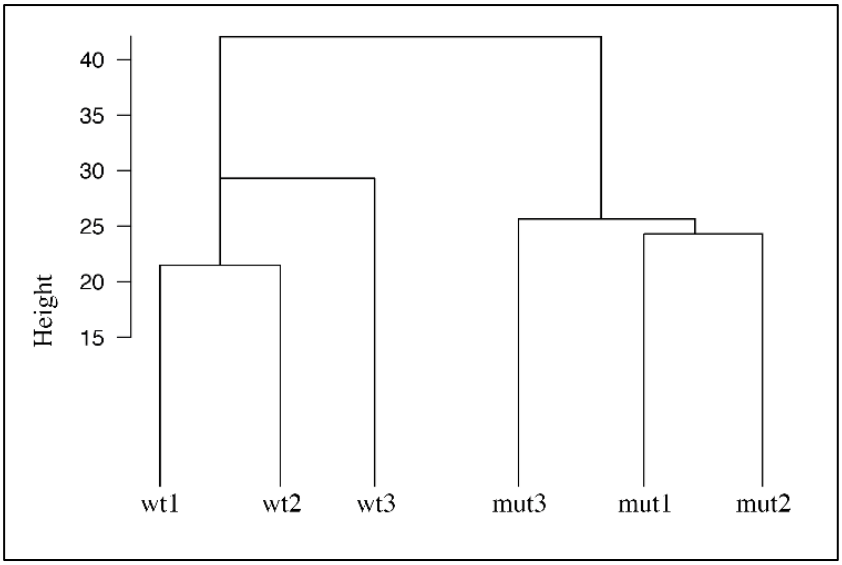


Figure 5.18 Dendrogram of rlog-transformed read counts for 6 samples. Pearson's correlation distance was computed between samples. The 2 conditions WT and mutant are separated.

Clustering analysis displayed as a heatmap of samples distances provided a clearly visible overview over similarities and dissimilarities between the replicates (Figure 5.19).

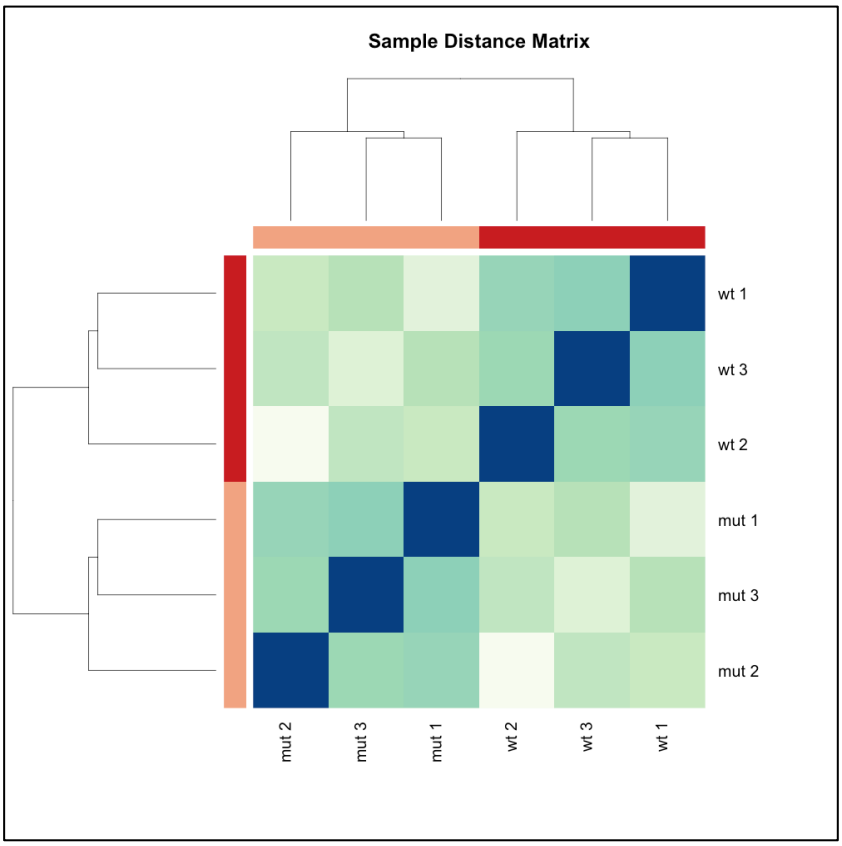


Figure 5.19 Examples of distance heat map. Filtered and normalised data were used for calculation of distances and drawing a distance heat map. Each square represents the correlation to all samples (including itself) with white for the lowest (0) and blue for the highest observed distance. The heatmap with clustering dendrograms showed a tight cluster for the 2 conditions WT (light pink box) and mutant (red box).

Connected to the distance matrix was the first principal components analysis (PCA) which is another way of examining the experiment variability and visualising the overall effect of experimental covariates and batch effects (Figure 5.20). From the PCA plot, I assessed whether the differences between the replicates (plotted in different shapes) were less strong than the differences due to condition associated with the mutant genotype (plotted in different colour). PCA analysis showed similar results as demonstrated by gene clustering analysis.

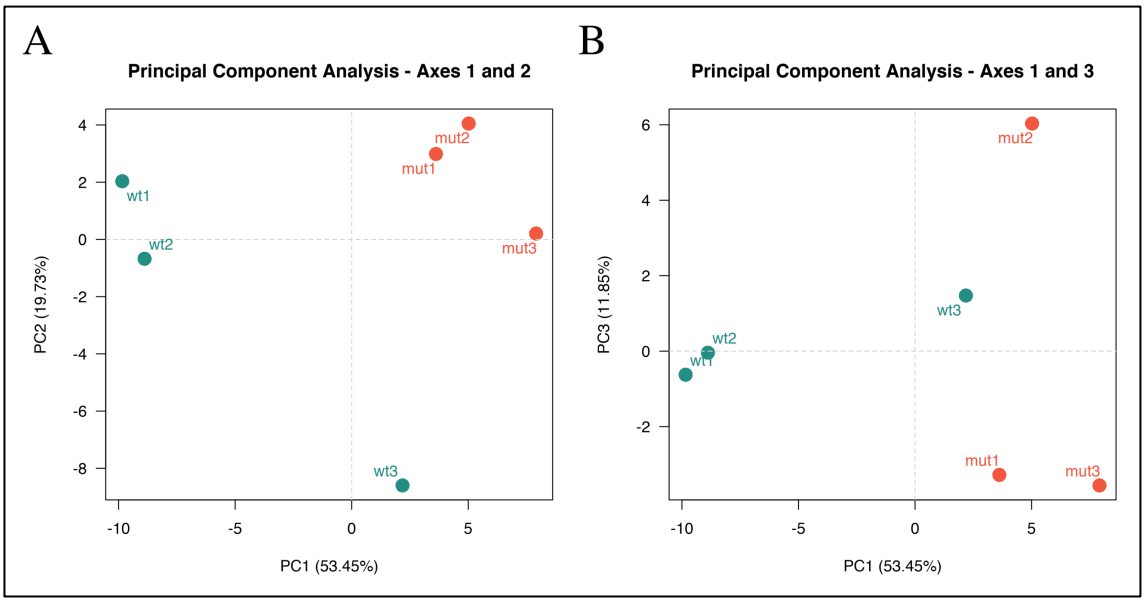


Figure 5.20 Example of PCA plot. PCA emphasises variation between samples and within conditions to identify strong patterns and clusters in the dataset. (**A**) First 2 components of the principal component analysis, with percentages of variance associated with each axis. PC1 is expected to separate samples from the different biological conditions (colour – WT green and mut red) and not by replicates; meaning that the biological variability and not differences between the samples within each batch was the main source of variance in the data. (B) First and third components of the principal component analysis. Both A and B highlight how sample wt3 shared more similar features to mutant condition than wt1 and wt2 samples capturing small biological fluctuations.

The following step was to identify the differentially expressed genes (DEGs) between the 2 conditions (WT vs mutant and GFP^- vs GFP^+). A variety of methods for differential gene expression analysis and software packages have been proposed over the years, including edgeR (Robinson et al., 2009), limma (Smyth, 2005), baySeq (Hardcastle and Kelly, 2010), NOIseq (Tarazona et al., 2015) and EBSeq (Leng et al., 2013). Benchmark tests comparing performance of different statistical methods has been assessed before (Costa-Silva et al., 2017; Seyednasrollah et al., 2015). In summary, number of biological replicates, sensitivity of the algorithm and type of library preparation are factors that influenced the analysis outcome (Schurch et al., 2016). Differential expression of my datasets was determined using the DESeq package (Love et al., 2014), this model works by fitting a linear model to account for the behaviour of the gene in each sample and analyse how gene expression varied due to the condition type. Considering the number of replicates, size and sequencing depth of my libraries, DESeq2 was preferable for differential expression analysis (Lamarre et al., 2018; Zhang et al., 2014).

The simplest design was to ask whether the DEGs could be explained only by the difference in genotype at a specific timepoint and to quantity genes as 'counts: ~ condition' where the condition was (WT vs mutant and GFP^- vs GFP^+). When I analysed the *sox32*$^{-/-}$ dataset at 5.25 hpf, I also introduce a 'batch' variable to account for major source of technical variations that affected the measurements in that specific data. The new design of the experiment was '~ batch + condition'. In all formulas, the most important factor (condition) was at the end to give it more 'weight'.

Finally, when I was interest in comparing developmental stages, I added the developmental time variable and then expanded the model by including an interaction term to determine how gene expression dynamics over developmental time was influenced by the genotype. The regression model for each gene *i* was updated as follows:

$$\begin{aligned}\text{Effect}_i = {} & a_i Condition_{status} + b_i Time_{gastrulation\ stage} \\ & + c_i\left(Condition_{status} : Time_{gastrulation\ stage}\right)\end{aligned}$$

meaning that I could test for the effect of genotype (condition) controlling for the effect of different time points and batch (the day when the samples were collected).

The DESeq2 algorithm estimated the models' coefficients and these coefficients were tested to get p-values and adjusted p-values. Figure 5.21 displays the distributions of raw p-values computed by the statistical test. The expected distribution should have a peak around 0 associated to the differentially expressed genes.

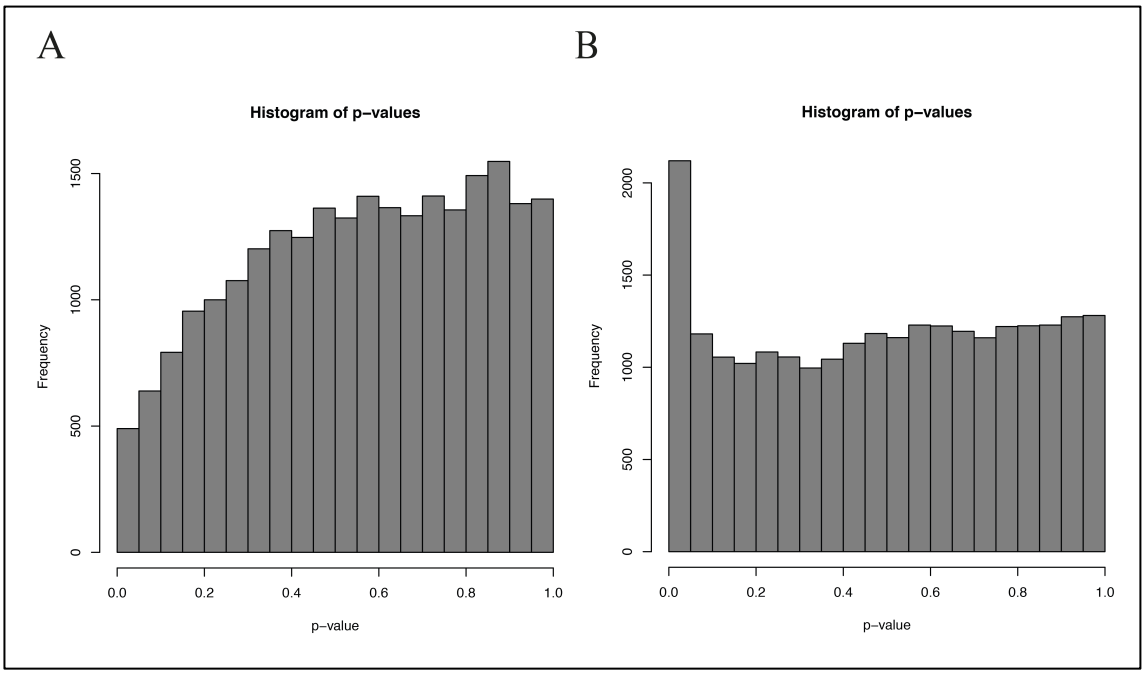


Figure 5.21 Example of distribution of raw p-values. A histogram of *p*-values is a graphical method to visualise whether the statistical test applied to the experiment is appropriate. The shape of the histogram distribution helps to identify potential problems with the statistical test, in particular if it is not appropriate. If the hypothesis being tested does not have the expected null distribution, the computed *p*-values are not valid and another statistical analysis pipeline needs to be applied. (**A**) Histogram of *p*-values for all genes tested for no difference between the 2 conditions, mutant and WT. No single peak was observed implying a problem with the model in describing gene expression in the dataset. This was *sox32*$^{-/-}$ 5.25 hpf before correcting for batch effect. (**B**) Flat distribution of *p*-values of *sox32*$^{-/-}$ 5.25 hpf after correcting for batch effect which were uniformly distributed between 0 and 1. Only 1 peak close to 0 represented the alternative hypothesis *p*-values – e.g. the difference between the conditions (mutant and WT).

A *p*-value adjustment was performed to take into account multiple testing and control the false positive rate. For this analysis, a BH *p*-value adjustment was used and the level of controlled false positive rate was set to 0.05 or 0.01 depending on the dataset.

After interpreting the histogram of p-values, I next visualised the differential expression analysis with MA and a volcano plot which is a clear and simple way to assess the results of the analysis (Figure 5.22). They show the relation between fold change and statistical confidence, differentially expressed transcripts that are at least 3-fold differentially expressed at a significance of 0.05 in any of the sample comparisons.

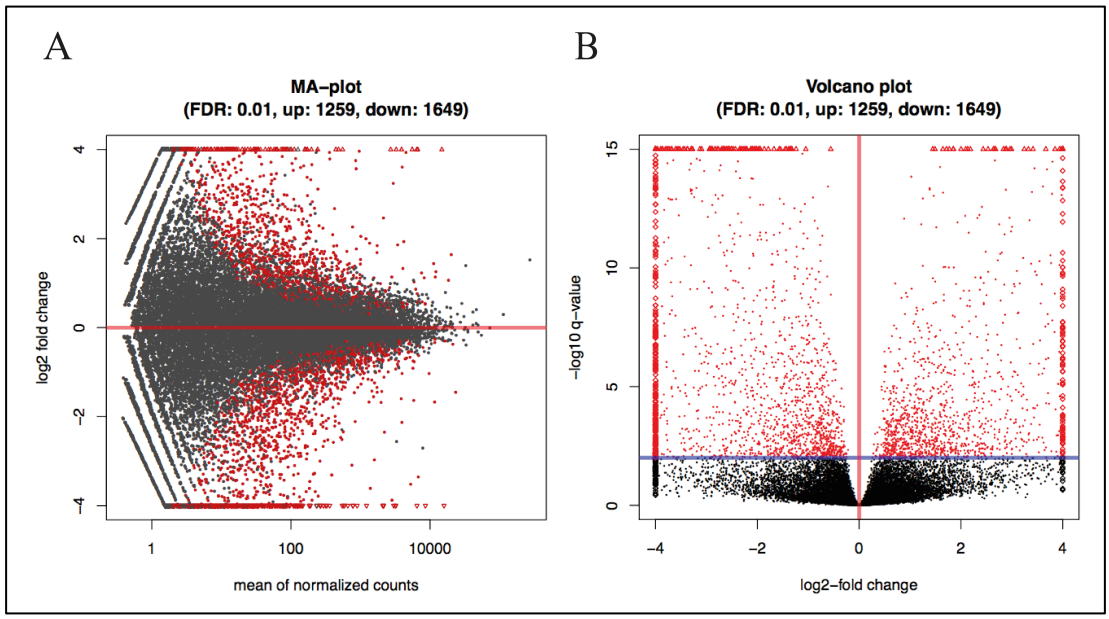


Figure 5.22 Quality control plots for differential gene analysis. (**A**) MA-plot. Differential gene expression depicted as MA-plot for the Mutant vs WT contrast at 9.00 hpf for *sox32*$^{-/-}$ dataset. The plot shows the relationship between the expression change (M) and the average expression strength (A), log ratio of differential expression as a function of the mean intensity for each gene. Red dots represent significantly differentially expressed genes. (**B**) Volcano plots for the comparisons performed and differentially expressed features are still highlighted in red. A volcano plot represents the log of the adjusted P value as a function of the log ratio of differential expression. Horizontal blue dotted lines indicate significance threshold ($p < 0.05$) whereas vertical lines indicate fold-change threshold (> 1-fold). Each circle represents one gene and all genes present on the analysis were plotted. Red points represent significantly differentially expressed gene by these criteria. Positive x-values represent upregulation and negative x-values represent downregulation. Triangles correspond to features having a too low/high FC to be displayed on the plot.

DESeq2 results produced a list of several hundred genes, this is because disruptions and perturbations to biological systems may affect large numbers of genes as the whole system becomes destabilised. After I ranked the lists of DEG by both p-values and fold change these gene were used both for building a heatmap and enrichment analysis. Both approaches were useful for detecting genes that were commonly regulated, or biological signatures associated with a particular condition (mutant genotype and endodermal cells transcriptome).

To display the gene expression data, the heatmap was combined with clustering methods in which groups of genes were joint based on the similarity of their gene expression pattern. The plot is a grid where each column is a sample and each row is a gene ranked by a specific feature (*p*-value or fold change). The intensity of the colour represents changes of gene expression. In my thesis, green represented upregulated genes and blue represented downregulated genes (Figure 5.23).

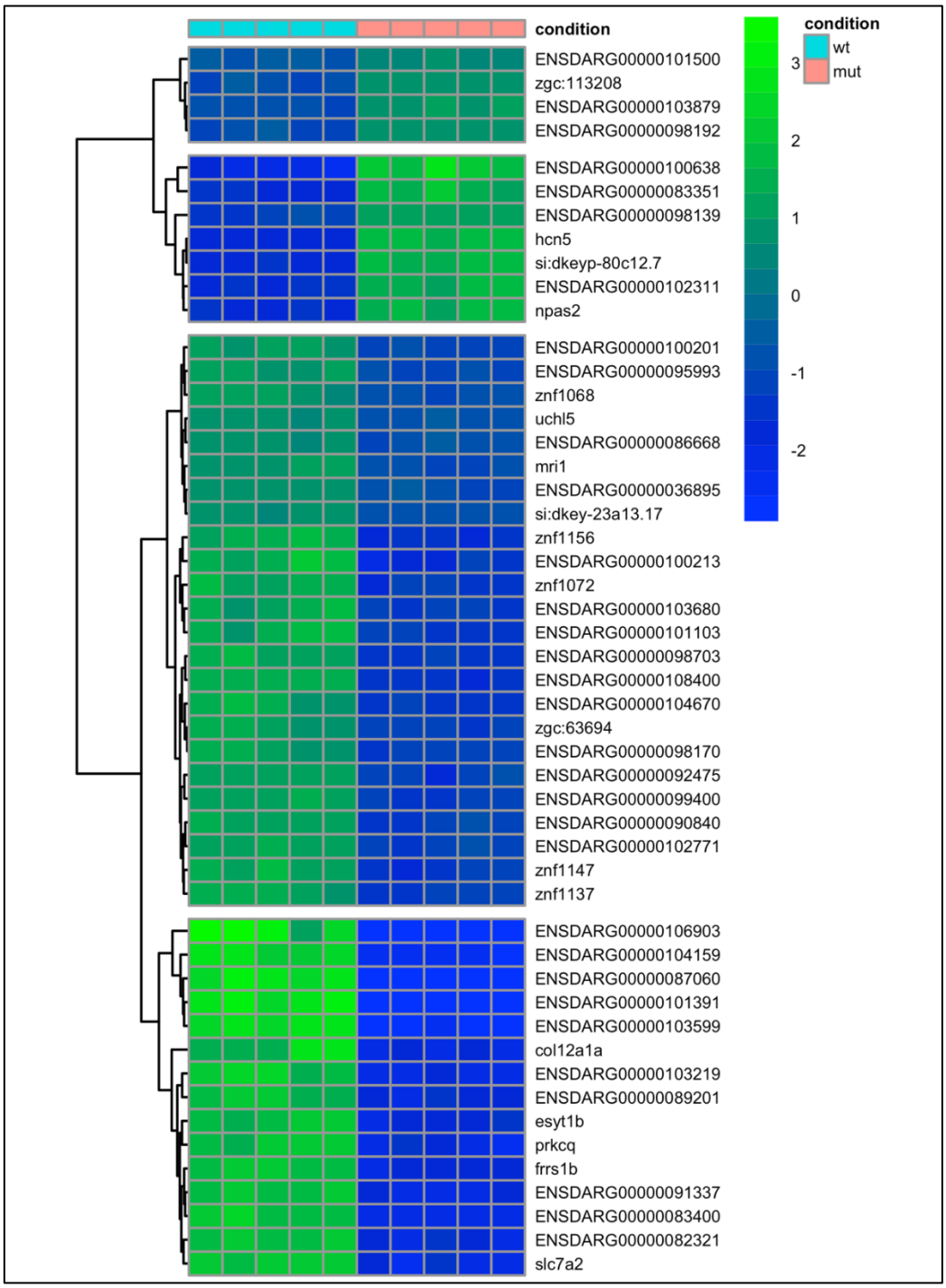
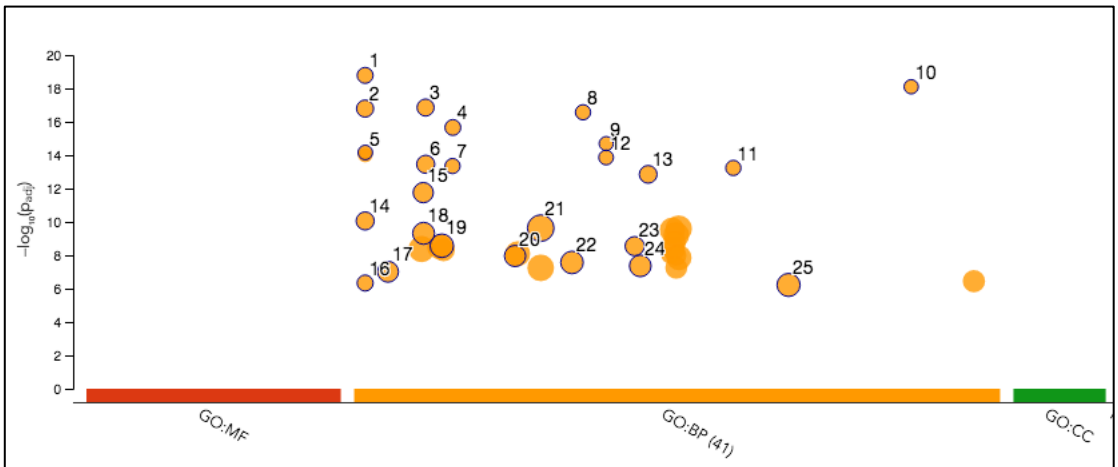


Figure 5.23 Examples of heatmap. The genes corresponding to the top 50 most differentially expressed genes in *mixl1* mutant were used to build the hierarchical clustering heatmap. Blue indicates low expression and green high. Genes sorted according to hierarchical cluster and scaled per row. This analysis clearly separated the groups of genes that were differentially expressed in the mutant vs. WT embryos.

The genes corresponding with the largest variance, both positive and negative co-efficient values in the model, were also selected for enrichment analysis and were submitted for signature analysis using 3 tools: g-profiler (Peterson et al., 2016) (Figure 5.24), PHANTER (Muruganujan et al., 2018) and ZEOGS (Prykhozhij et al., 2013)(Table 5.3).



id	source	term id	term name	p_{adj} (query_1)
1	GO:BP	GO:0001706	endoderm formation	1.663×10^{-19}
2	GO:BP	GO:0001704	formation of primary germ layer	1.614×10^{-17}
3	GO:BP	GO:0007492	endoderm development	1.403×10^{-17}
4	GO:BP	GO:0010470	regulation of gastrulation	2.200×10^{-16}
5	GO:BP	GO:0001714	endodermal cell fate specification	6.895×10^{-15}
6	GO:BP	GO:0007498	mesoderm development	3.520×10^{-14}
7	GO:BP	GO:0010453	regulation of cell fate commitment	4.497×10^{-14}
8	GO:BP	GO:0035987	endodermal cell differentiation	2.692×10^{-17}
9	GO:BP	GO:0042663	regulation of endodermal cell fate specification	2.036×10^{-15}
10	GO:BP	GO:1903224	regulation of endodermal cell differentiation	7.969×10^{-19}
11	GO:BP	GO:0060795	cell fate commitment involved in formation of p...	5.845×10^{-14}
12	GO:BP	GO:0042659	regulation of cell fate specification	1.379×10^{-14}
13	GO:BP	GO:0045995	regulation of embryonic development	1.414×10^{-13}
14	GO:BP	GO:0001708	cell fate specification	8.769×10^{-11}
15	GO:BP	GO:0007369	gastrulation	1.761×10^{-12}
16	GO:BP	GO:0001707	mesoderm formation	4.626×10^{-7}
17	GO:BP	GO:0003002	regionalization	9.697×10^{-8}
18	GO:BP	GO:0007389	pattern specification process	4.851×10^{-10}
19	GO:BP	GO:0009790	embryo development	2.682×10^{-9}
20	GO:BP	GO:0022603	regulation of anatomical structure morphogene...	1.106×10^{-8}
21	GO:BP	GO:0032502	developmental process	2.365×10^{-10}
22	GO:BP	GO:0035295	tube development	2.697×10^{-8}
23	GO:BP	GO:0045165	cell fate commitment	2.842×10^{-9}
24	GO:BP	GO:0045595	regulation of cell differentiation	4.324×10^{-8}
25	GO:BP	GO:0072359	circulatory system development	6.001×10^{-7}

Figure 5.24 Example of g-profiler plot. A Manhattan plot that illustrated the enrichment analysis results for endodermal genes. In the scatterplot, each circle represents an enriched term. Functional terms are colour-coded and ranked on the x-axe (red: molecular function, orange: biological process, green: cellular component, blue: transcription factor). The y-axis indicated the adjusted enrichment p-values in $-\log_{10}$ scale. The light circles represent significant terms which are reported in the bottom table.

Table 5.3 Example of ZEOGS results. The output of ZEOGS was a list of significant anatomical term where the input genes was enriched. ZEOGS used information available in ZFIN databases.

Anatomical term	P-value
E-YSL	**0.005**
presumptive paraxial mesoderm	**0.036**
margin	**0.031**

presumptive endoderm	**0.039**
presumptive brain	0.139
YSL	0.168
organizer inducing center	0.145
I-YSL	0.176
DEL	0.181
forerunner cell group	0.374
blastoderm	0.343
presumptive blood	0.350
presumptive mesoderm	0.381
yolk	0.480
anatomical structure	0.818
axis	0.927

In summary, the RNA-seq processing pipeline was used to extract knowledge from my study that profiled gene expression using RNA-seq in zebrafish comparing 2 conditions, I now describe each dataset in detail.

5.5 *mixl1*$^{-/-}$ transcriptome

In this section, I present RNA-seq results for *mixl*$^{-/-}$ embryos. In zebrafish development, Mixl1 regulates the generation of mesoendodermal precursor cells. In the *mixl1*$^{-/-}$, cells expressing *sox32* and *sox17* are reduced in number compared to wildtype embryos and proper development of the gut tube and the heart is hindered, with mutants displaying a cardia bifida phenotype similar to what is observed in *sox32*$^{-/-}$ (Kikuchi et al., 2000). Thus, the identification genes downstream of *mixl1* was a step towards elucidating its role in mesendodermal bifurcation, in particular to dissect why, for example, the number of *sox17* expressing cells is decreased in the mutants. What are the changes in the core of the mesendodermal bifurcation kernel when Mixl1 protein is not functional and are we able to identify alternative and novel markers that act in this network? We know that mesendodermal precursors become physically separated during gastrulation (van Boxtel et al., 2015; van Boxtel et al., 2018) but how is this reflected at the genetic level?

Transcriptome analysis is often used to identify differentially expressed genes that may underpin unique biological properties of cells. The transcriptome of *mixl1* deficient embryos presented in this study was used to identify genes that were regulated by Mixl1 protein at the beginning of zebrafish gastrulation. Mapping phenotypes to genotype changes has been one of the long-standing aims in biology and performing transcriptome analysis has accelerated

and simplified the tackling of this problem. In order to identify genes regulated by Mixl1, I used a mutant line generated in Didier Y.R. Stainier's lab and proceeded with a classic mutant vs WT comparison (Kikuchi et al., 2000). *mixl1*$^{-/-}$ embryos are known to have fewer endodermal cells, however no full genomic transcriptomic profile has been done. RNA preserved in TRIzol (3 WT replicates + 3 mutant replicates) was collected and shipped from D.M. lab in Innsbruck Austria; in addition, 10 homozygote adults were also kindly shipped. These adults were generated from embryos of a heterozygous crosses that were rescued by injecting *mixl1* RNA at the 1-cell stage, then genotyped and homozygotes grown to adulthood; the rescue experiments were necessary because this specific locus mutation is lethal in homozygote zebrafish.

As reported previously, the quality of extracted RNA from the shipped samples yielded a RIN < 6, I therefore decided to i) use a ribosomal depletion strategy for the library preparation and ii) collect additional RNA triplicates from new embryos produced by the adult homozygotes fish at KCL (RIN > 9). In total for this experiment therefore, 6 WT libraries (3 Austria + 3 KCL) and 6 *mixl1*$^{-/-}$ libraries (3 Austria + 3 KCL) were sequenced and analysed. All 12 libraries were prepared using the same method – ribosomal depletion for consistency.

5.5.1 Ribosomal RNA depletion

As mentioned earlier, rRNA depletion uses probes that selectively bind rRNA sequences to capture these molecules and efficiently remove them from the sample. Most of the commercial kits available are tested in human, mouse and rat and not zebrafish. Firstly, I tested NEBNext rRNA Depletion Kit which was inconsistent in hybridizing to rRNA and subsequently in the removal of rRNA from the samples (Figure 5.25).

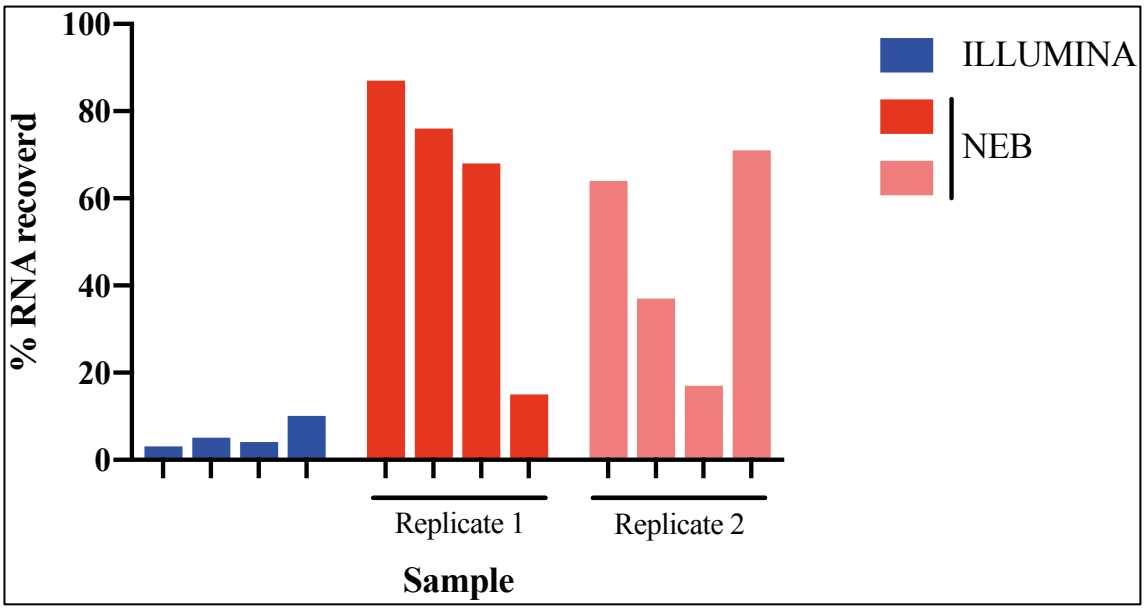


Figure 5.25 rRNA removal pilot test. Each column was a sample. The experiment was repeated twice using NEB kit (shades of red) with inconsistent results. Illumina kit (in blue) was used to directly prepare *mixl1*$^{-/-}$ libraries. As described in literature, expected recovery rate was around 10%.

In the literature, previous researchers (Lee et al., 2013; Nudelman et al., 2018; Trinh et al., 2017) have used the RiboZero rRNA removal kit from Illumina to characterise changes in the transcriptome associated with the ZGA transition in zebrafish development; the Illumina kit, although expensive, was able to capture extremely abundant rRNAs efficiently and removed 80–90% of the total RNA samples (Figure 5.25), while the NEB kit did not consistently capture rRNAs. Efficient removal of rRNA is critical to enable cost-effective sequencing of RNA samples, I therefore used this strategy to prepare all the *mixl1*$^{-/-}$ libraries.

5.5.2 Read alignment and quality control

Next, I sought to map the reads. The total number of samples was 12, with 6 WT samples and 6 *mixl1*$^{-/-}$ samples. The reads in the FASTQ files were aligned to the zebrafish genome with STAR and loaded in R to analyse differential gene expression using DESeq2. A summary of the library characteristics is provided in Table 5.4. Read depth was comparable among all samples.

Table 5.4 Summary of total reads for *mixl1*$^{-/-}$ libraries.

Austria	WT			Mutant		
	Replicate 1	Replicate 2	Replicate 3	Replicate 1	Replicate 2	Replicate 3
Total reads	11.4 M	11.2 M	13.0 M	10.6 M	11.8 M	10.1 M
KCL	WT			Mutant		
	Replicate 1	Replicate 2	Replicate 3	Replicate 1	Replicate 2	Replicate 3
Total reads	10.6 M	14.9 M	14.4 M	14.6 M	11.9 M	10.9 M

I decided to first explore the dataset assuming no differences in the libraries (All) and then I also analysed the Austria samples (Austria) and KCL samples (KCL) separately. In order to determine the relative similarity of the replicates I used PCA. PCA projects multidimensional data, the counts for each of the genes in the transcriptome, onto a 2 dimensional space keeping the relative distances between points as much as possible. Close points share high similarity in the transcriptome. The resulting plots in Figure 5.26 exhibited strong clustering of the replicates for the conditions in all analyses (All, Austria and KCL), except for 2 outlier samples (WT2 Austria and *mixl1*$^{-/-}$*1*). Furthermore, for all data sets, the dominant principal component (PC1) was linked to the genotype, thus the dominant component of variability of my data associated directly with the variables of interest (mutant genotype). As shown in Panel A, PCA revealed tight clustering of 5 mutant replicates independently from the batch (Austria or KCL) and the libraries were highly reproducible as the data points were clustered together with small variability. More variability was observed for WT replicates along the PC2. In particular Panel C (KCL samples) showed replicate WT1 was further away from the other 2 WT samples in the graph, suggesting the gene expression profiles of the WT1 samples were distinct from the other 2 WT samples. However, this separation was along PC2 which was not associated with the effects of the mutant genotype.

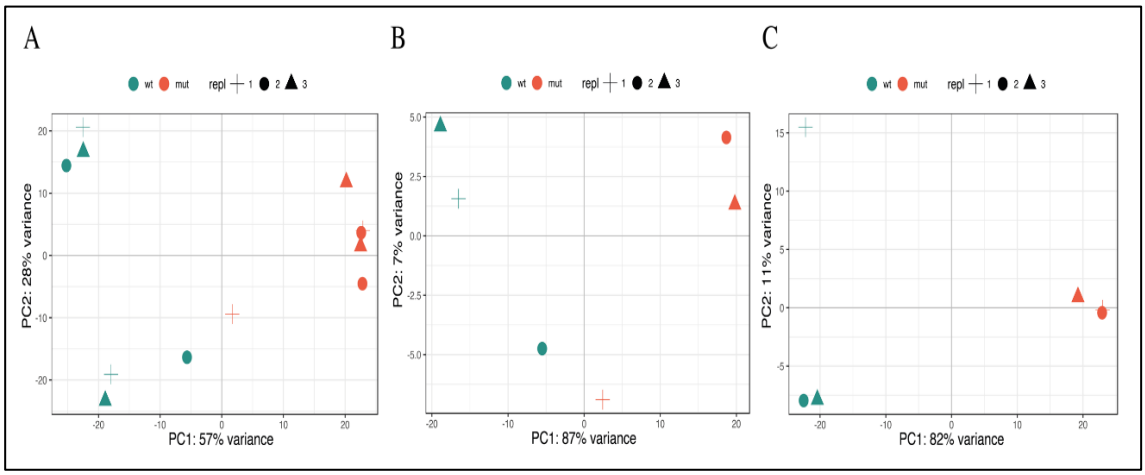


Figure 5.26 *mixl1*$^{-/-}$ PCA plot. WT (green) and *mixl1*$^{-/-}$ (red) were separated by a large distance, suggesting that their gene expression profiles were quite different, thus the main variability within the experiment originated from biological differences between the conditions and not the samples. (**A**) Analysis of all 12 samples. WT replicate 2 and mutant replicate 1 were the closest samples. (**B**) Only Austria samples and (**C**) KCL samples. PC1 linked to the mutant genotype explained 57% of the variability in (A) 87% in (B) and 82% in (C), indicating that altered gene expression patterns were primarily linked to *mixl1* mutation.

The goal of this study was to identify gene regulatory interactions associated with Mixl1, therefore I then assessed differential gene expression using the DESeq2 package.

5.5.3 Differential expression analysis 5.25 hpf

When I compared all 12 libraries, a total of 4071 genes exhibited significant changes in expression levels (FDR < 0.01, absolute $\log_2$FC > 1), a majority (2075/4071 or 51%) of the DEGs were downregulated in the mutant, including known *mixl1* signature genes such as *foxa2* and *mixl1* (Nelson et al., 2017). On the other hand, 1996 were upregulated in the *mixl1*$^{-/-}$.

Multiple differentially expressed genes were not associated with a known gene name or anatomical term(s) on ZFIN, and manual inspection of the genes suggests many were likely to encode non-coding RNAs. While these genes are likely to be important in the normal development of mesendodermal cells, they were harder to classify and interpret.

Differential expression analysis on the Austria libraries revealed 584 significantly more abundant and 534 less abundant genes in the *mixl1* mutant transcriptome compared to WT. For the KCL libraries, DESeq2 reported 1065 significantly downregulated and 1436 upregulated in the *mixl1* transcriptome vs WT transcriptome; all the analyses were completed

with the same parameter at FDR < 0.01. Summaries of all the differential genes expressed are reported in Table 5.5 and visualised with volcano plots in Figure 5.27.

Table 5.5 Total number of DEGs for *mixl1*$^{-/-}$.

Contrast	Total No. of Significant DEGs	No. of Upregulated Genes	No. of Downregulated Genes
mixl1$^{-/-}$ (6) vs. WT (6)	4071	1996	2075
mixl1$^{-/-}$ (3) vs. WT (3) - Austria	1118	584	534
mixl1$^{-/-}$ (3) vs. WT (3) - KCL	2501	1436	1065

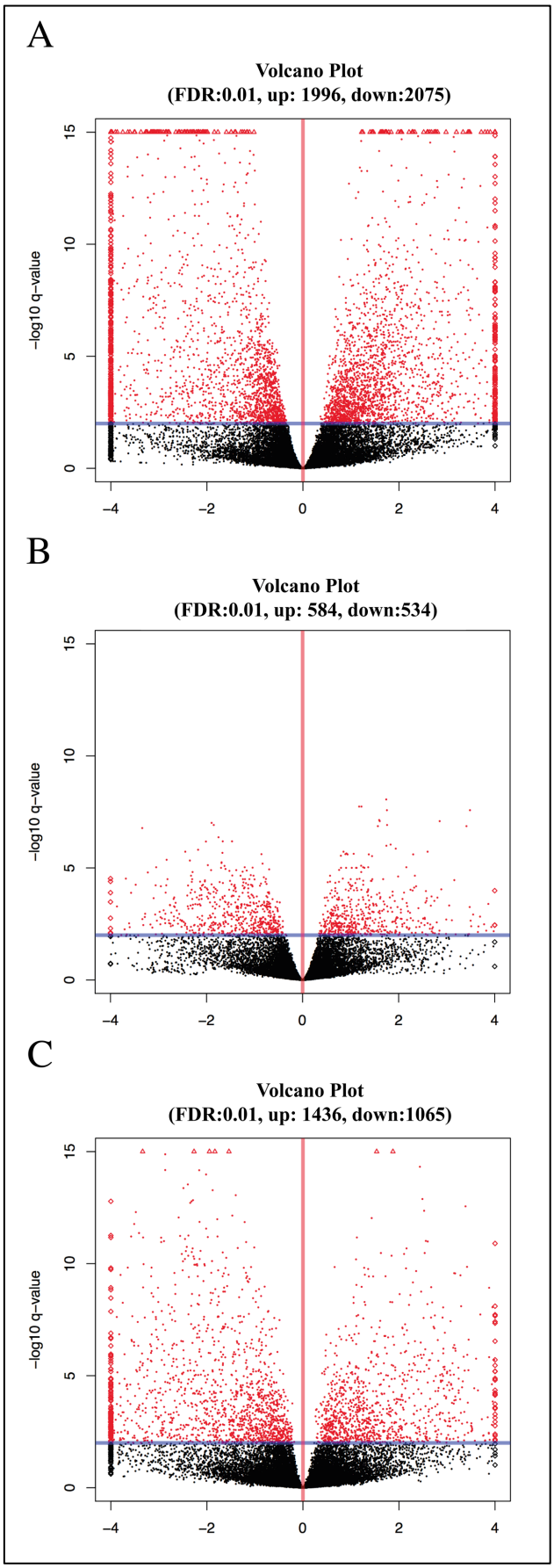


Figure 5.27 Volcano plot for *mixl1*$^{-/-}$ datasets. A large set of genes were differentially expressed in *mixl1*$^{-/-}$ in comparison to WT embryos. (**A**) Analysis for all 12 libraries, (**B**) Austria samples and (**C**) KCL samples.

Differentially expressed genes were categorized as those having at least 1-fold difference between the 2 conditions with statistical significance (FDR < 0.01) and are reported in red. Blue horizontal line is the FDR threshold.

Expression profiles were also visualised using a heatmap (Figure 5.28), in which both the replicates (columns) and expression levels of significant DEGs (rows) were ordered by hierarchical clustering to reveal their hidden structure. Use of a colour coding scheme for upregulated and downregulated expression helped in the recognition of groups of genes with concordant expression patterns across the condition, and assessment of the heterogeneity of the samples.

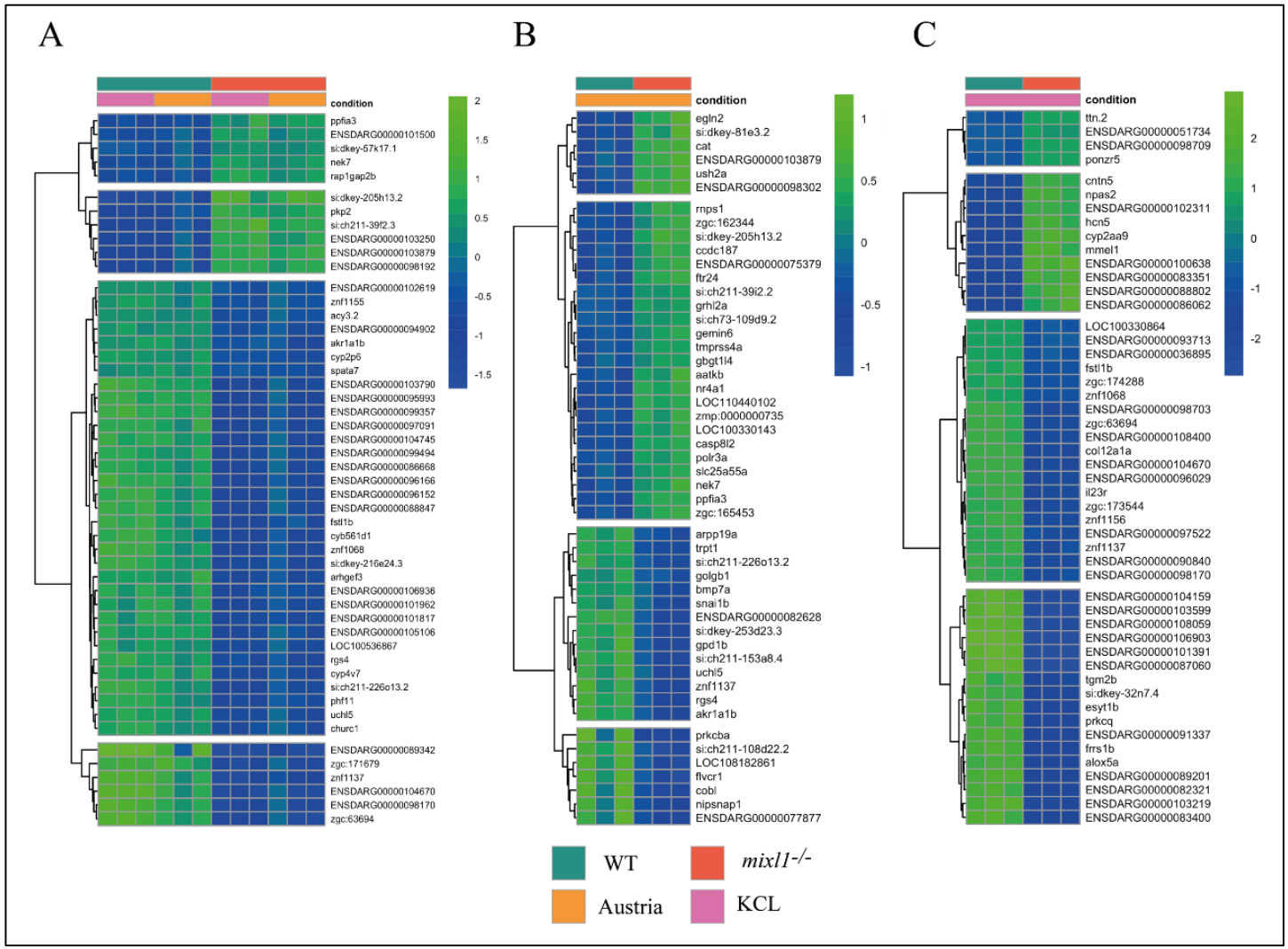


Figure 5.28 Heatmap visualisation and hierarchical clustering of mixl1$^{-/-}$ expression data. Heatmap of top 50 upregulated DEGs in All (A), Austria (B) and KCL (C) demonstrated unambiguous divergence of gene expression in the 2 groups. Note that the majority of genes were reported with the Ensembl Gene ID(s) and not with a gene name; most of DEGs in this dataset were uncharacterised genes, possibly non-coding. Log_2-fold enrichment is presented in blue-green colour key.

Next, I questioned whether the differential analysis was dependent on which datasets I was using, in particular, I asked whether the DEGs were similar in the 3 aforementioned analyses. I compared the upregulated and downregulated genes for All, Austria and KCL, which shared 222 downregulated and 305 upregulated genes (Figure 5.29).

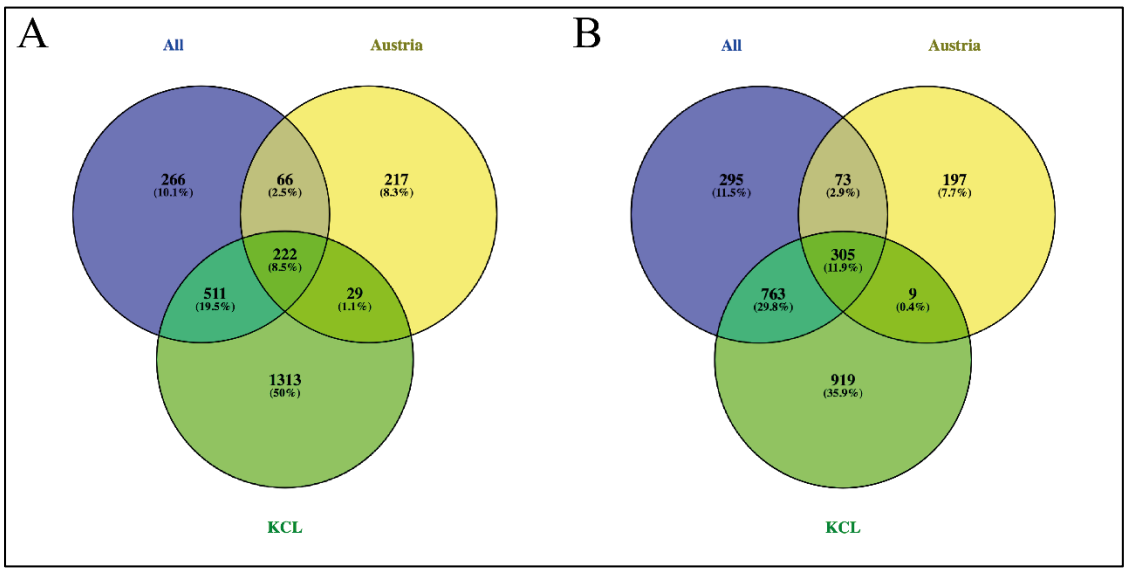


Figure 5.29 Intersectional analysis of DEGs from *mixl1*$^{-/-}$ identified by RNA-seq. Numbers of significantly differentially expressed genes (> 1-fold change and FDR < 0.01) in the 3 separate analysis were identified. The Venn diagrams analysis showed (**A**) only 222 common downregulated genes, which were present in all 3 different types of analysis whereas (**B**) 305 common genes were consistently upregulated.

The 222 top downregulated genes included *mixl1, foxa2, lft1, dusp27, notum1a, ddit4, prdx1, notum1a, smad5, ddit4 and irx3a*.

Autoregulation is a known mechanism exploited to control genetic pathways and it has evolved to refine and tightly maintain TF levels. It has been argued that master developmental regulators, which control large numbers of genes, employ autoregulation to strictly control their levels of expression, therefore lower level of *mixl1* transcripts in *mixl*$^{-/-}$ where not surprising (Crews and Pearson, 2009; Hermsen et al., 2010). *Mixl1* occurrence correlates with the developmental importance of determining mesendodermal cell fates hence *mixl1* autoregulation may be a valuable property to fine-tuning the mesendodermal regulatory circuits. Lower levels of *foxa2* in *mixl*$^{-/-}$ were reported in the paper describing the *mixl*$^{-/-}$ mutant phenotype (Kikuchi et al., 2000), whereas *lft1* and *dusp27* were interesting downstream candidate genes.

Lft1 is a well characterised Nodal antagonist and *dusp4*, which has an inhibitory role downstream of Fgf signalling, is important in endoderm specification (Brown et al., 2008). Recent studies have revealed the interplay between Nodal agonist, Nodal antagonist, Fgf signalling and mitogen-activated kinase (MAPK) signalling cascades in dampening and balancing the Nodal domain in the margin at the beginning of gastrulation which allows endoderm to form close to the margin (Brown et al., 2008; van Boxtel et al., 2018).

Given the known roles of *dusp4* during endoderm development, it was possible that *dusp27* which is a phylogenetic relative, could also play a role in regulating endodermal intracellular signalling pathway. *dusp27* mutants display impaired movements and despite being able to go through somitogenesis, maturation of the contractile apparatus in myofibers is altered. The molecular pathways through which *dusp27* acts and its function in endoderm formation at present are unknow (Rogers et al., 2017).

Loss-of-function studies revealed that *notum1a* depletion results in limited expansion of Wnt/β-catenin signalling through interaction of specific GPI-anchored proteins such as Gpc3. Importantly, *notum1a* has no effect on induction of Tgdf1, a GPI-anchored EGF-CFC cofactor required for proper Nodal signalling and *notum1a* overexpression does not alter *sox32*-expressing cells or *gsc*-expressing cells (Flowers et al., 2012). Thus, Notum1a is a Wnt specific deacylase and understanding the relationship of *mixl1* and *notum1a* could be of great value in the study of signalling feedback between Nodal, Wnt and β-catenin pathways in the presumptive mesendoderm territory.

Smad1/Smad5 are activated by BMP signalling which inhibit Nodals and regulates endoderm formation; in addition, *smad5* mediates endodermal pouch morphogenesis and subsequent craniofacial development (Lovely et al., 2016; Poulain et al., 2006).

Moreover, *prex1* is another Nodal target which regulates endodermal cell motility (Woo et al., 2012). All in all, the above identified target genes are direct and indirect components of Nodal signalling and are of interest for future studies to expand Nodal signalling and the kernel of genes that establish mesendodermal cells fate.

Finally, *irx3a* was identified in the endoderm modules downstream of *gata5* (Tseng et al., 2011). Irx3a is a homeobox transcription factor like Mixl1 and play a central role in endocrine pancreas development (Pauls et al., 2007).

Taken together these results suggest that Mixl1 initiates specifying endodermal fate through known endodermal genes and an array of novel putative genes and play a key part in establishing the necessary and sufficient triggers for initiating development of genes expressed in late endodermal and mesodermal structures. In addition, this approach associated Mixl1 to a putative novel role in regulating member of other signalling pathway besides Nodal.

5.5.4 Enrichment analysis

Functional enrichment analysis was performed using 3 online tools ZEOGS, g:profiler and PANTHER to interpret the biological functions. For the analysis I considered the upregulated and downregulated genes separately. The downregulated genes were ordered based on FDR and then used as input. Only functional categories containing more than 3 genes were included in the analysis and I used Benjamini-Hochberg FDR at 0.05 for correcting for multiple testing and to adjust significance thresholds.

The results from ZEOGS confirmed that *mixl1* regulates multiple genes associated with endoderm (*p-adj*= 0.039) and paraxial mesoderm *(p-adj*= 0.036). The genes are also mapped to the margin (*p-adj*= 0.031) and YSL (*p-adj*= 0.005) which are the location of origin of mesendodermal cells (Table 5.6).

Table 5.6 ZEOGSS results for the enrichment of top upregulated common genes *mixl*$^{-/-}$

Anatomical term	P-value
YSL	**0.005**
presumptive paraxial mesoderm	**0.036**
margin	**0.031**
presumptive endoderm	**0.039**
presumptive brain	0.139
E-YSL	0.168
organizer inducing centre	0.145
I-YSL	0.176
DEL	0.181
forerunner cell group	0.374
blastoderm	0.343
presumptive blood	0.350
presumptive mesoderm	0.381
yolk	0.480
anatomical structure	0.818

axis	0.927

However, both g:profiler and PANTHER failed to visualise the many-to-many relationships between GO terms and annotate function specific driver genes for the 222 gene list; this was not surprising considering that most of the DEGs (90 out of 222) were unannotated in ZFIN as pointed out by ZEOGSS (Table 5.7).

Table 5.7 ZEOGGS output showing the following 90 genes did not have anatomical terms on ZFIN.

Gene name	Description
rab43	RAB43, member RAS oncogene family
adpgk2	ADP-dependent glucokinase 2
exoc1	exocyst complex component 1
si:ch1073-190k2.1	si:ch1073-190k2.1
trim35-30	tripartite motif containing 35-30
dap1b	death associated protein 1b
si:ch211-107e6.5	si:ch211-107e6.5
tmem59l	transmembrane protein 59-like
ppm1h	protein phosphatase, Mg2+/Mn2+ dependent, 1H
cry5	cryptochrome 5
si:ch211-257p13.3	si:ch211-257p13.3
kirrel3a	kin of IRRE like 3 a
prdx1	peroxiredoxin 1
si:ch211-223a21.3	si:ch211-223a21.3
c1galt1c1	C1GALT1-specific chaperone 1
si:dkeyp-86c4.1	si:dkeyp-86c4.1
si:dkey-17o15.2	si:dkey-17o15.2
grin2db	glutamate receptor, ionotropic, N-methyl D-aspartate 2D, b
cacna2d1a	calcium channel, voltage-dependent, alpha 2/delta subunit 1a
zgc:77086	zgc:77086
ibtk	inhibitor of Bruton agammaglobulinemia tyrosine kinase
si:dkeyp-2e4.3	si:dkeyp-2e4.3
mri1	methylthioribose-1-phosphate isomerase 1
si:dkey-271j15.3	si:dkey-271j15.3
tmem101	transmembrane protein 101
si:dkey-24p1.6	si:dkey-24p1.6
sypl2a	synaptophysin-like 2a
si:dkey-73p2.2	si:dkey-73p2.2
arnt2	aryl-hydrocarbon receptor nuclear translocator 2
zgc:171566	zgc:171566
btr12	bloodthirsty-related gene family, member 12
nipsnap1	nipsnap homolog 1 (C. elegans)
zdhhc3a	zinc finger, DHHC-type containing 3a

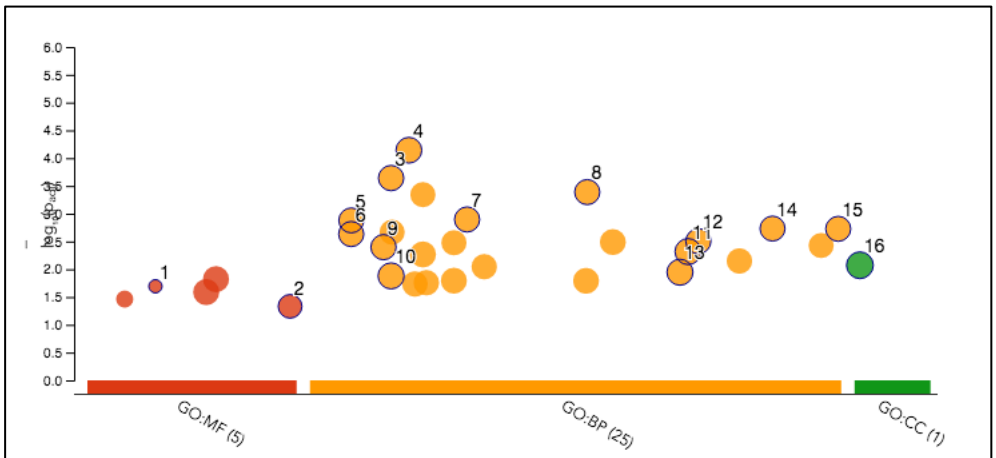
sult2st3	sulfotransferase family 2, cytosolic sulfotransferase 3
hmgn2	high mobility group nucleosomal binding domain 2
si:dkey-201g16.1	si:dkey-201g16.1
zgc:113295	zgc:113295
usf1l	upstream transcription factor 1, like
hist1h4l	histone 1, H4, like
arpp19a	cAMP-regulated phosphoprotein 19a
wtip	Wilms tumor 1 interacting protein
si:ch211-197e7.3	si:ch211-197e7.3
acy3.2	aspartoacylase (aminocyclase) 3, tandem duplicate 2
si:dkey-247i3.6	si:dkey-247i3.6
uchl5	ubiquitin carboxyl-terminal hydrolase L5
odf3b	outer dense fiber of sperm tails 3B
ppp2r5ca	protein phosphatase 2, regulatory subunit B', gamma a
mzt2b	mitotic spindle organizing protein 2B
tpi1b	triosephosphate isomerase 1b
si:dkey-20i20.8	si:dkey-20i20.8
cps1	carbamoyl-phosphate synthase 1, mitochondrial
s100s	S100 calcium binding protein S
si:dkeyp-93d12.1	si:dkeyp-93d12.1
rgs4	regulator of G-protein signaling 4
si:ch211-110e21.4	si:ch211-110e21.4
zgc:171679	zgc:171679
zgc:174268	zgc:174268
si:ch211-113a14.22	si:ch211-113a14.22
si:dkey-6d5.1	si:dkey-6d5.1
uros	uroporphyrinogen III synthase
si:ch73-347e22.8	si:ch73-347e22.8
xylb	xylulokinase homolog (H. influenzae)
zgc:173545	zgc:173545
ldhba	lactate dehydrogenase Ba
il23r	interleukin 23 receptor
vars	valyl-tRNA synthetase
wu:fc75a09	wu:fc75a09
exosc3	exosome component 3
cbr1l	carbonyl reductase 1-like
si:ch211-108d22.2	si:ch211-108d22.2
fxyd6l	FXYD domain containing ion transport regulator 6 like
rbb4l	retinoblastoma binding protein 4, like
cyr61l2	cysteine-rich, angiogenic inducer, 61 like 2
zgc:174224	zgc:174224
si:dkey-32n7.4	si:dkey-32n7.4

cyp4v7	cytochrome P450, family 4, subfamily V, polypeptide 7
akr1a1b	aldo-keto reductase family 1, member A1b (aldehyde reductase)
si:dkey-273g18.1	si:dkey-273g18.1
tmem138	transmembrane protein 138
si:ch211-130m23.2	si:ch211-130m23.2
si:dkey-4e4.1	si:dkey-4e4.1
hebp2	heme binding protein 2
atp6v1d	ATPase, H+ transporting, lysosomal V1 subunit D
plscr3b	phospholipid scramblase 3b
si:ch1073-513e17.1	si:ch1073-513e17.1
phlda1	pleckstrin homology-like domain, family A, member 1
ncam3	neural cell adhesion molecule 3
zgc:174696	zgc:174696
ttc27	tetratricopeptide repeat domain 27
ncam2	neural cell adhesion molecule 2

I then followed up the enrichment analysis by manually eliminating the unannotated genes from the list and resubmitting to PANTHER and g:profiler.

PANTHER recognised an enrichment for a family of zinc finger proteins (C2H2 zinc finger transcription factor (*p-adj*=1.21E^{-06}) and zinc finger transcription factors (*p-adj* =6.99E^{-05})). This was of particular interest knowing that *mixl1* and *gata5*, both zinc-finger genes, regulate *sox32* expression and that White et al. 2017 identified a set of related zinc finger proteins highly active from 4.30 to 8.00 hpf.

Furthermore, the g:profiler results highlighted the interconnection of these genes with the regulation of biological processes, regulation of RNA synthesis and DNA binding transcription factors, suggesting Mixl1 functions in controlling expression of other transcriptional regulators in endoderm and mesoderm lineages during gastrulation (Figure 5.30).



id	source	term id	term name	p_{adj} (query_1)
1	GO:MF	GO:0016922	nuclear receptor binding	2.030×10^{-2}
2	GO:MF	GO:0140110	transcription regulator activity	4.658×10^{-2}
3	GO:BP	GO:0010468	regulation of gene expression	2.287×10^{-4}
4	GO:BP	GO:0016070	RNA metabolic process	7.216×10^{-5}
5	GO:BP	GO:0006355	regulation of transcription, DNA-templated	1.323×10^{-3}
6	GO:BP	GO:0006351	transcription, DNA-templated	2.329×10^{-3}
7	GO:BP	GO:0032774	RNA biosynthetic process	1.273×10^{-3}
8	GO:BP	GO:0051252	regulation of RNA metabolic process	4.096×10^{-4}
9	GO:BP	GO:0009889	regulation of biosynthetic process	4.009×10^{-3}
10	GO:BP	GO:0010467	gene expression	1.320×10^{-2}
11	GO:BP	GO:0090304	nucleic acid metabolic process	4.818×10^{-3}
12	GO:BP	GO:0097659	nucleic acid-templated transcription	3.135×10^{-3}
13	GO:BP	GO:0080090	regulation of primary metabolic process	1.133×10^{-2}
14	GO:BP	GO:1903506	regulation of nucleic acid-templated transcripti...	1.856×10^{-3}
15	GO:BP	GO:2001141	regulation of RNA biosynthetic process	1.877×10^{-3}
16	GO:CC	GO:0005634	nucleus	8.484×10^{-3}
17	KEGG	KEGG:03018	RNA degradation	4.990×10^{-2}

Figure 5.30 Enrichment analysis for *mixl1*$^{-/-}$. Manhattan plot representing statistical overrepresentation analysis by g:profiler of the upregulated gene list. GO terms related to nuclear receptor binding and transcription regulator activity shown in red. The analysis showed the presence of signalling pathways implicated in RNA synthesis, general development and DNA-protein and RNA-protein interactions.

In addition, the multiple unannotated genes in the DEGs list raise a potential function of *mixl1* in orchestrating a set of non-coding RNAs yet to be annotated and characterised (genes with only Ensembl IDs).

In conclusion, the zebrafish *mixl1* transcriptome at 5.25 hpf was in agreement with existing knowledge on genes expressed in mesendodermal cells at this developmental stage and raises potential avenues for further investigation. Primarily, it supports a subset of genes that are coexpressed at the beginning of gastrulation and then later in development that are expressed exclusively in endodermal and mesodermal cells, thus capturing the intricate interaction between these 2 germ layers.

5.5.5 RT-qPCR validation

The results of the enrichment analysis correlated *mixl1*$^{-/-}$ DEGs with YSL, margin and endodermal/mesodermal genes, therefore I proceeded to confirm some of these targets using RT-qPCR. I made this choice as my ultimate goal was to update the GRN around endoderm formation, establishing new connections around Mixl1. The following genes were selected to validate their predicted expression by RT-qPCR: *foxa2*, *mixl1*, *lft1*, *irx3a*, *smad5*, *notum1a* and *ddit4*. All 6 were predicted to be downregulated in *mixl1* mutant. The fold changes of expression to WT were $\log_2$ transformed and plotted (Figure 5.31). The RT-qPCR results showed that the expression patterns of 5 out of these 6 genes were consistent between RT-qPCR and the RNA-seq data, the only exception was *notum1a* (p = 0.07*)*, which was not significantly downregulated. These results suggested that the RNA-seq exploratory predictions on the *mixl1* mutant transcriptome were reliable. As shown in Figure 5.31, the expression of *foxa2*, *mixl1* and *smad5* were validated to be the most highly downregulated in the absence of Mixl1 protein regulatory function. Mixl1 controls its own transcript levels, endodermal genes (*foxa2*) and mesodermal genes (*irx3a, notum1a*, *ddit4)* and Nodal associated genes (*lft1*, *smad5).*

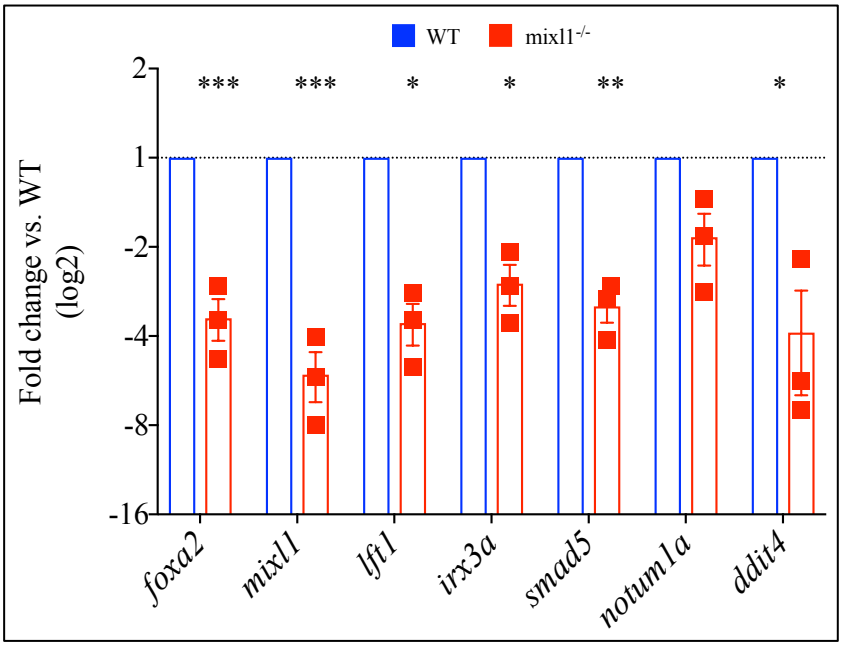


Figure 5.31 RT-qPCR validation of *mixl1*$^{-/-}$ RNA-seq data. Fold change of RT-qPCR data on seven genes showing downregulated expression in the mutant embryos. Data were obtained from 3 independent biological replicates, normalised to *ef1α* and expression was calculated relative to WT (blue bars) and log2 transformed. Mean and SEM are shown; unpaired two-tailed t-test, *$p \leq 0.05$, **$p \leq 0.01$, ***$p \leq 0.001$.

Taken together, these results suggested that *mixl1* could play a role in fine tuning the extent mesendodermal domain at the margin and may also have a crucial mechanical role during endodermal and mesodermal cell migration at the onset of gastrulation.

5.6 *sox32*$^{-/-}$ transcriptome

A described in Chapters 1 and 3, Sox32 is important for endoderm specification (Alexander et al., 1999; Dickmeis et al., 2001; Kikuchi et al., 2001). Identifying downstream targets of Sox32 should therefore help to pinpoint genes involved in endoderm development at different stages and outline how the role of Sox32 evolves during gastrulation, including the associated changes in downstream gene regulation. Sox32 plays an important role in regulating the expression of a subset of endodermal genes during developmental cell-type specification in different modules. Endodermal cells differentiate into different cell types, possibly due to Sox32 acting differently on gene cohorts depending on specific cues.

5.6.1 Read alignment and quality control

High-throughput RNA-seq was performed on 6 libraries per time point, 5.25 hpf and 9.00 hpf (3 *sox32*$^{+/+}$ samples and 3 *sox32*$^{-/-}$ samples). As described previously, read quality was checked using FastQC, reads were aligned to the Ensembl zebrafish reference genome using STAR and --quantMode parameter was used to extrapolate counting reads per gene. A summary of library characteristics is provided in Table 5.8. Read depth was comparable among the conditions and between the 2 time points.

Table 5.8 Summary of total read for *sox32*$^{-/-}$ libraries.

5.25 hpf	WT			Mutant		
	Replicate 1	Replicate 2	Replicate 3	Replicate 1	Replicate 2	Replicate 3
Total reads	15.8 M	16.9 M	17.9 M	20.1 M	16.5 M	17.6 M
9.00 hpf	WT			Mutant		
	Replicate 1	Replicate 2	Replicate 3	Replicate 1	Replicate 2	Replicate 3
Total reads	17.5	16.7 M	15.8 M	20.4 M	13.3 M	15.3 M

PCA was performed to detect outliers and any batch effects for both 5.25 hpf (Figure 5.32) and 9.00 hpf datasets (Figure 5.33). The ComBat function in the 'sva package' (Leek JT, 2019) was implemented on the dataset at 5.25 hpf to adjust for batch effect. DESeq2 was then applied to perform normalisation and test for differential expression.

Figure 5.32 illustrates the improvement of the dataset after correcting for batch effect. As shown in panel A, samples were clustered together by the replicate variable rather than being associated to WT/mutant condition. After applying ComBat (panel B), a clear separation between WT samples on the left and the mutant samples on the right side was visible. PC1, the component associated with the genotype, explained 53% of the variance of the system. Samples heatmaps and hierarchical clustering were also employed; similarity between the samples was computed as Pearson correlation. The cluster was visualised both before and after batch effect correction (Panel C and D).

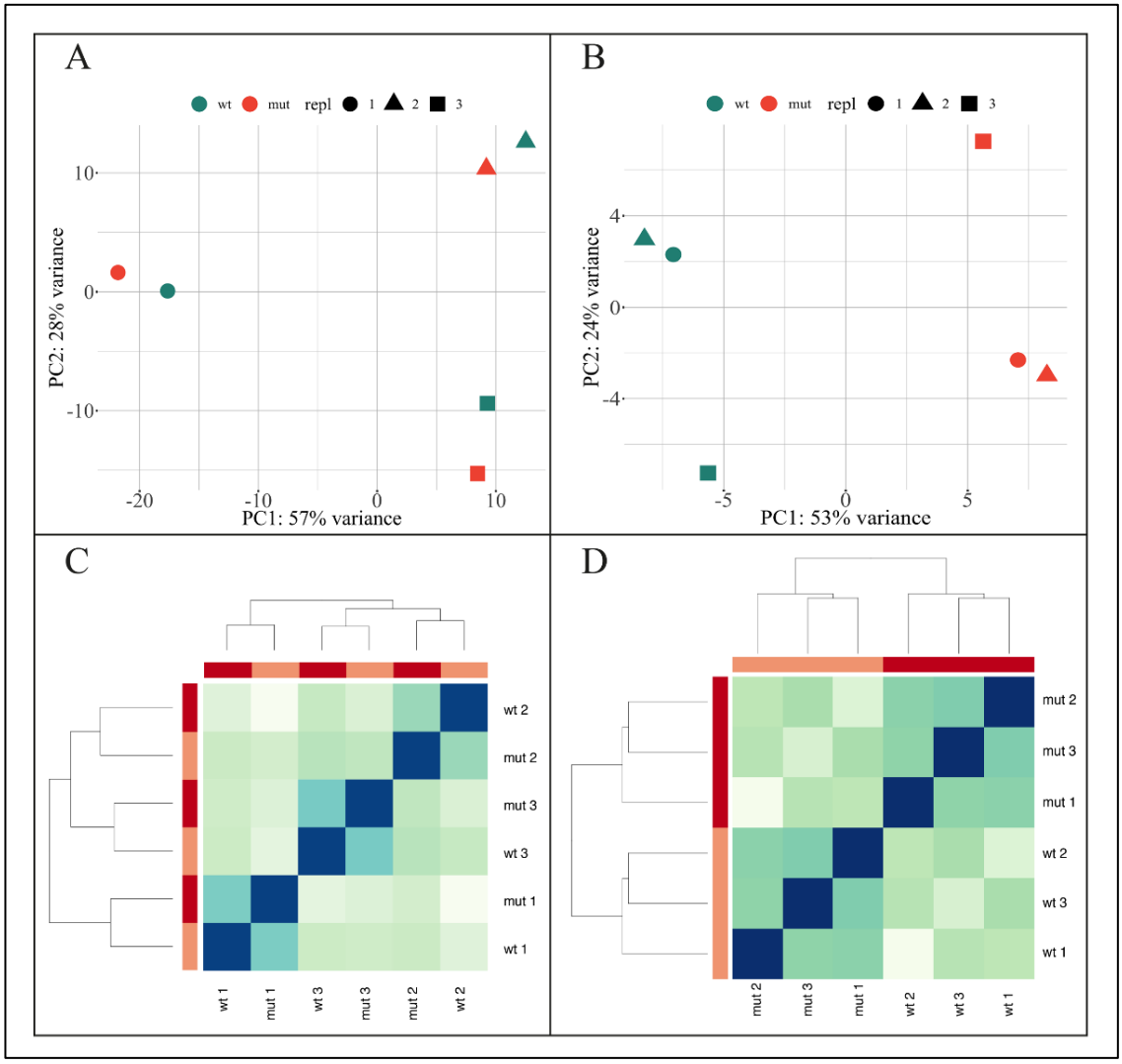


Figure 5.32 Batch effect in *sox32*$^{-/-}$ dataset at 5.25 hpf. PCA plot before (**A**) and after (**B**) removal of

the batch bias. The plot visually indicates the distances or dissimilarities between each dataset. The distance between the samples in the plot can be interpreted as the log fold change for the gene expression. (A) Strong batch effect for the replicate/date was observable in this dataset with close clustering of WT and mutant by replicate, rather than condition. (B) After using ComBat to account for the batch bias, noticeable improvement of the dataset was observable with the replicates clustering together whilst the samples from the 2 conditions were well separated. Hierarchical cluster and similarity matrix plots before (**C**) and after (**D**) correction for the replicate batch bias, using ComBat function. The colour scale indicates the degree of correlation (dark = high, light = low). The plots were generated by DESeq2, using the one minus Pearson correlation coefficient. For downstream analysis, the replicate effect was taken into account in the statistical models for DEGs in DESeq2.

Figure 5.33 shows the PCA plot and Pearson distance matrix for the RNA-seq libraries at 9.00 hpf. Both PCA and the correlation analysis found 2 clusters between the biological replicates, with PC1 explaining the 53% of the variability of the mutant genotype.

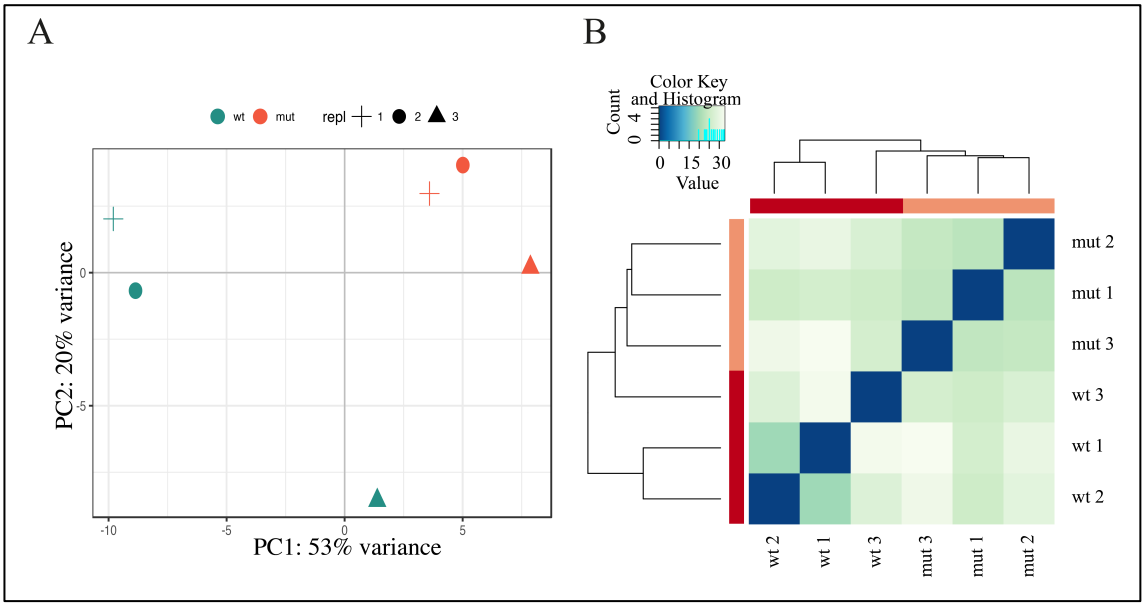


Figure 5.33 PCA plot and pairwise Pearson correlation coefficients in *sox32*$^{-/-}$ dataset at 9.00 hpf. (**A**) PCA plot of gene expression data showing the 1st and 2nd principal components. PCA showed a clear association with the genotype along the first axis (PC1) explaining most of the variance in the system, indicating that altered gene expression patterns were primarily linked to *sox32* mutation. (**B**) Correlation matrix of 6 RNA-seq libraries. Samples were hierarchically clustered with the Pearson correlation distance method for comparison among transcriptomes. The colour scale indicates the degree of correlation. The correlation matrix and heatmap were generated using R software.

Before proceeding to the next step in the bioinformatics pipeline I performed an additional quality-control on the 9.00 hpf dataset. Although the PC1 associated with the genotype explained 53% of the variability of the system in Figure 5.33A, it could be argued that WT

replicate 3 shared more attributes of PC1 with the mutant replicates than within its own group of WT replicates. To justify my downstream analyses, I temporarily excluded some replicates from the analysis. I tested both 2 vs 2 design (WT1,2 vs Mut1,2) and a 2 vs 3 design (WT1, 2 vs Mut1,2,3) and re-ran the comparison. As shown in Figure 5.34, in the 2 vs 2 analysis, PC1 now explained 66% of the variance associated with the mutant transcriptome, whereas in the 2 vs 3 analysis, PC1 explained 76% of the variance. However, when I performed the DEG analysis in DESeq2, the most upregulated and downregulated genes were not strongly affected by omitting the third replicate, suggesting that i) the DESeq2 model worked better with 6 samples to detect the DEGs and ii) the core of genes affected by the *sox32* mutation was robust, meaning the signature of the mutant was strongly different and detectable from the WT for a subset of genes. I therefore decided to keep all 6 samples.

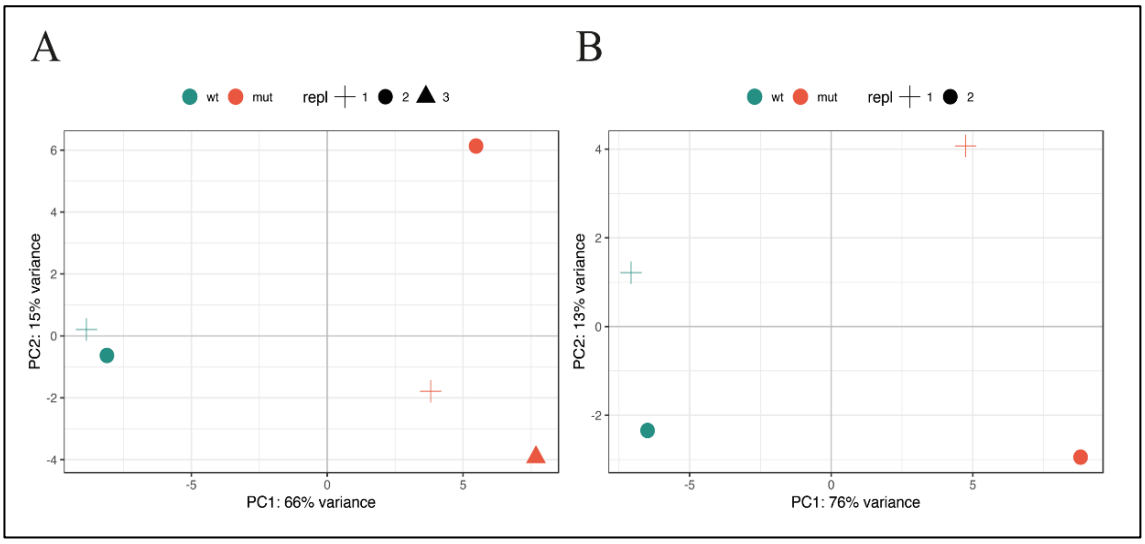


Figure 5.34 PCA plots in the *sox32* mutant dataset at 9.00 hpf. (**A**) PCA plot using only 2 WT and 3 mutant replicates. (**B**) PCA plot using only 2 WT and 2 mutant replicates.

5.6.2 Differential expression analysis 5.25 hpf

I first performed differential gene expression analysis on the 5.25 hpf datasets and as shown in the heatmap in Figure 5.35; only 8 genes were significantly differentially expressed (FDR < 0.1) when I did not account for batch effect. The reassuring result was to see how *sox17*, a known direct target of Sox32, was detected as significantly downregulated in this preliminary analysis.

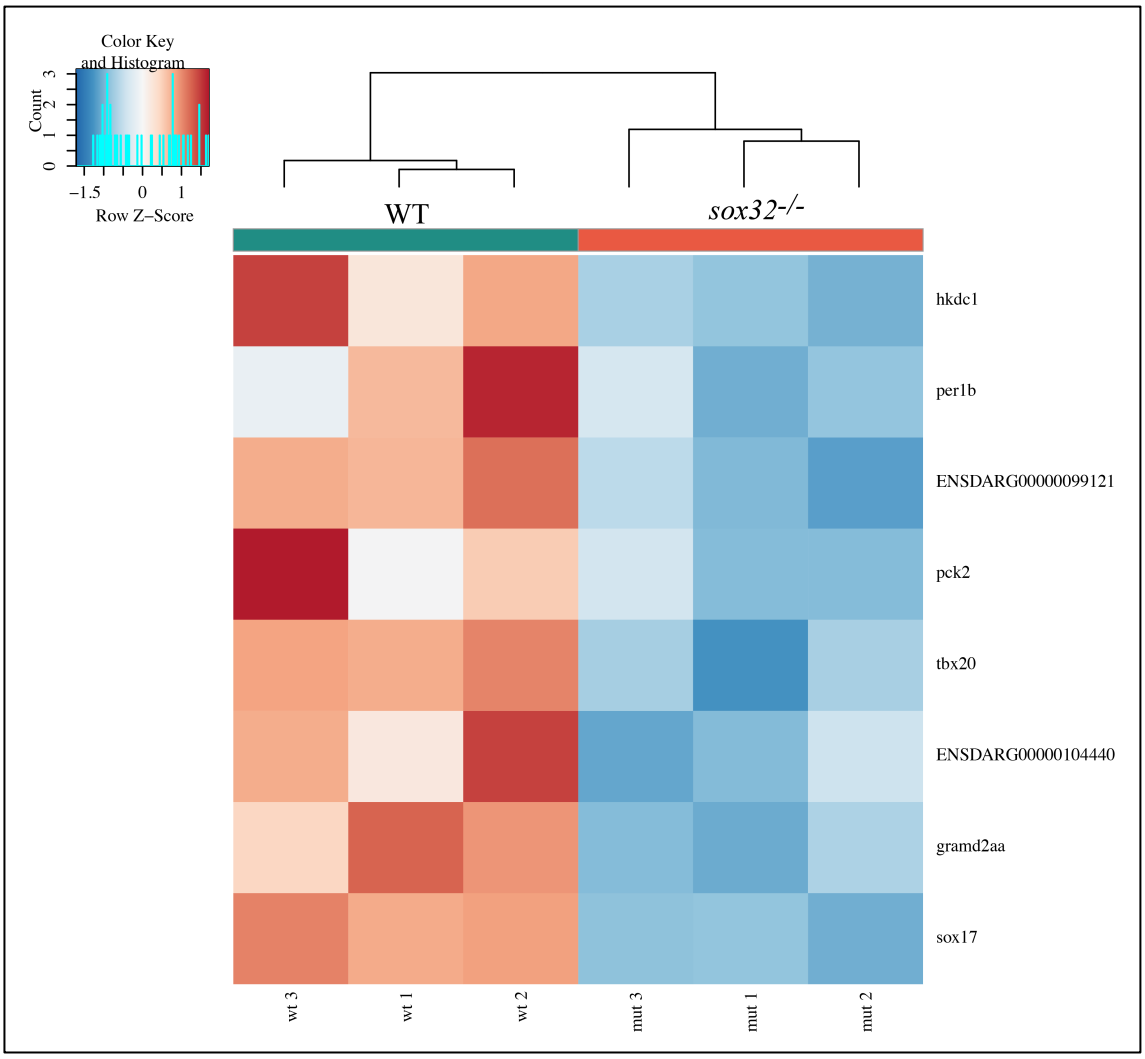


Figure 5.35 Heatmap of DEGs for *sox32*$^{-/-}$ at 5.25 without batch correction. Columns were samples, rows were genes; the standardized expression levels were depicted by colour gradient: upregulated genes in red, downregulated genes in blue. Only 8 DEGs were found by DESeq2 analysis using a model that did not incorporate a batch effect term. *sox17*, a known direct target of *sox32* was among the differentially expressed genes. The heatmap gave the pattern of expressional changes of the 8 significantly differentially expressed genes.

I then re-ran the analysis after using ComBat and incorporating a batch effect term in the DESeq2 model (from ~ condition to ~ batch + condition). Differential expression analysis revealed 444 genes differentially expressed ($\log_2$FC > 1 and FDR < 0.05) between mutant and WT samples, including 192 that were upregulated in the *sox32*$^{-/-}$ group and 253 that were downregulated (Figure 5.36).

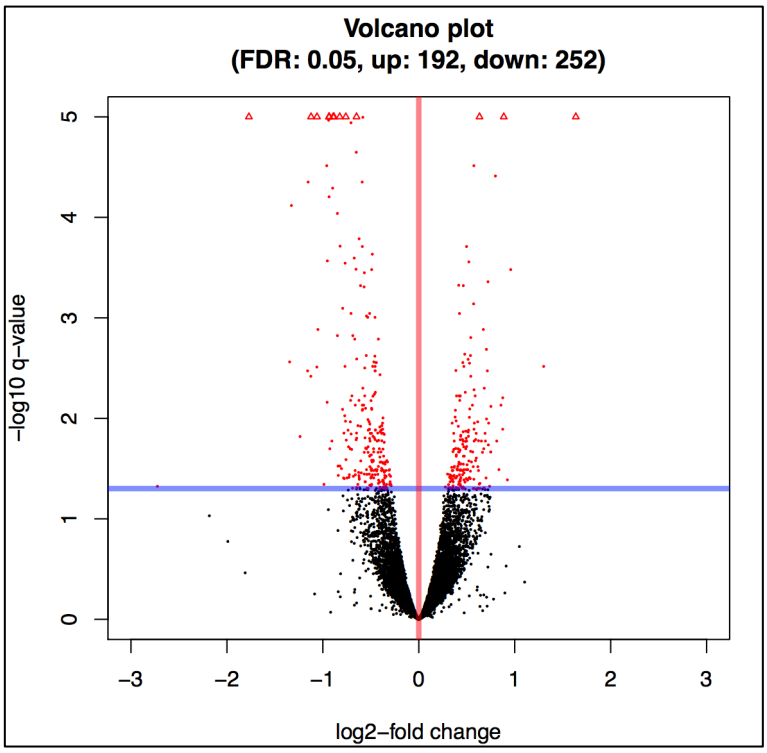


Figure 5.36 Volcano plot for *sox32*$^{-/-}$ at 5.25 hpf. Low expression genes (log_2TPM < 5) were excluded from the analysis. Significantly differentially expressed genes log_2FC >1 and FDR < 0.05 (blue line)) are shown in red; genes with no change in expression are shown in black.

The top downregulated genes included previously identified genes such as *sox17*, a known direct target of *sox32,* and genes which are required in the endoderm to pattern the pancreas and determine the pancreatic beta cells number such as *cdx4* (Kinkel et al., 2008). Amongst the downregulated genes were mesodermal markers such as *sp5l*, *dlc* and *tbxta* (Morley et al., 2009; Nelson et al., 2017). *cdh6* which is involved in nephrogenesis and promotes neural crest cell detachment was also among the downregulated genes (Clay and Halloran, 2014; Straub et al., 2011). In mouse development, Cdh6 expression is observed within the endodermal cell populations (Inoue et al., 2008), however *cdh6* function has not yet been describe in zebrafish endoderm. The top upregulated genes include *mxtx2* and *nanog*, 2 important TFs involved in activating Nodal signalling in the YSL (Xu et al., 2012). A visual representation of the top DEGs ranked by *p*-value is shown in Figure 5.37.

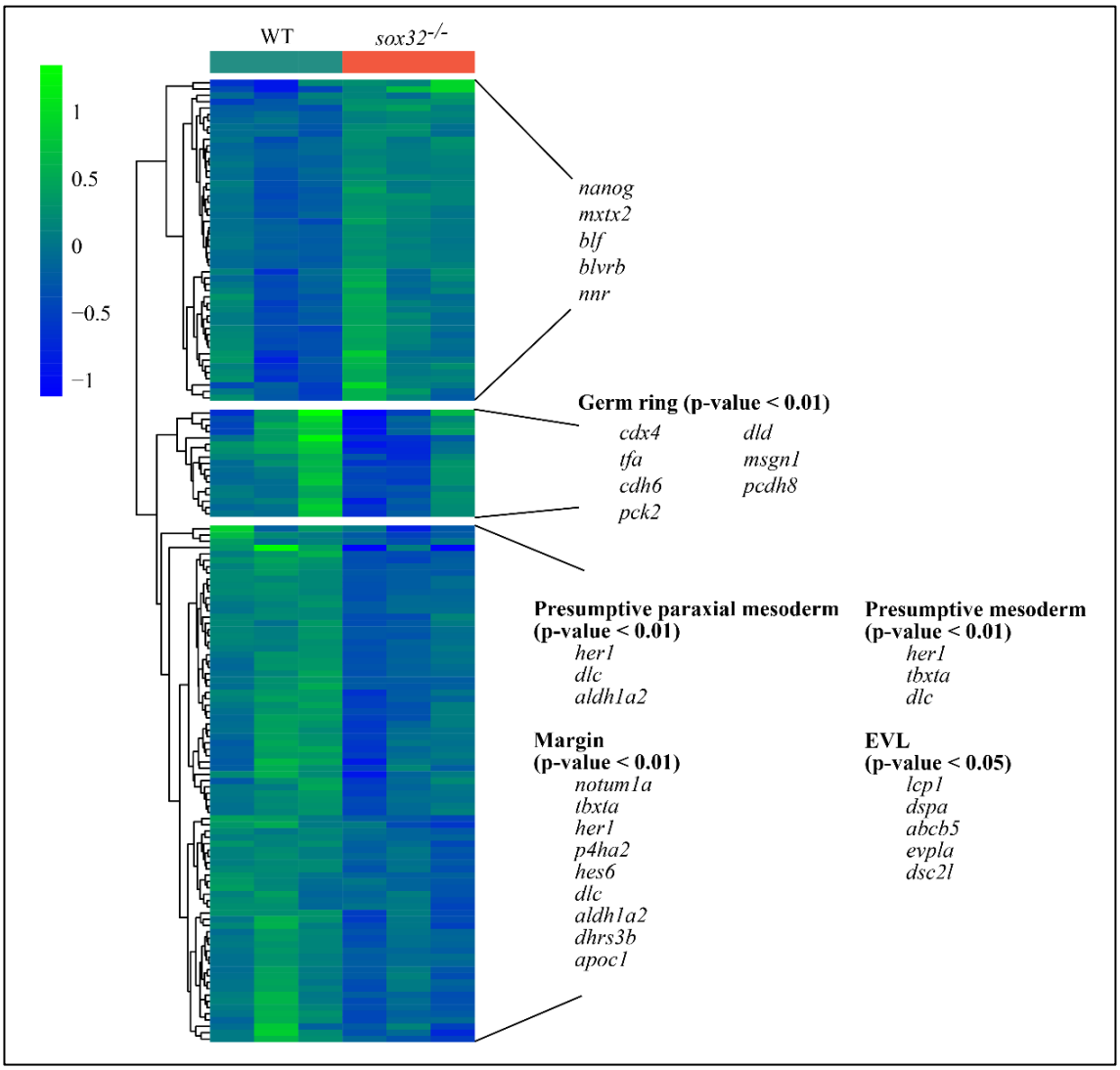


Figure 5.37 Heatmap summarising the top 150 DEGs in *sox32*$^{-/-}$ at 5.25 hpf. Genes were selected and classified into 3 groups. Examples of genes that belong to group 1 are *nanog* and *mxtx2*. Examples of genes that belong to group 2 were *cdx4*, *cdh6* and *pck2*. Mesodermal genes such as *dlc* and *tbxta* belonged to group 3. Log_2-fold enrichment is presented in blue-green colour key.

5.6.3 Enrichment analysis at 5.25 hpf

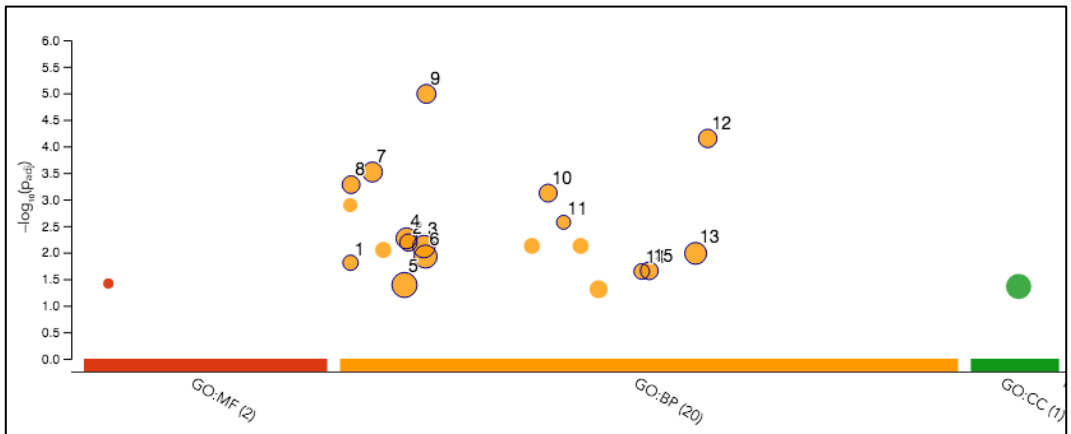
GO analysis was performed to gain deeper insights into the functions of the gene sets. ZEOGS summarised the expression pattern properties of gene sets as E-YSL (p-value= 0.005), presumptive paraxial mesoderm (p-value= 0.036), margin (p-value= 0.031) and presumptive endoderm (p-value= 0.039) (Table 5.9). The results implied that *sox32* is highly involved in regulating the cascade of genes acting in the dorsal margin of the embryo at the beginning of gastrulation and in regulating Nodal signalling in this area. It also implied that multiple

unknown genes in the list could be tested as novel markers for endodermal cell populations in further studies.

Table 5.9 ZEOGS enrichment analysis. Top 10 most enriched GO terms obtained from genes that were significantly more highly expressed in *sox32*$^{-/-}$ mutant.

Anatomical term	P-value
E-YSL	**0.005**
presumptive paraxial mesoderm	**0.036**
margin	**0.031**
presumptive endoderm	**0.039**
presumptive brain	0.139
YSL	0.168
organizer inducing center	0.145
I-YSL	0.176
DEL	0.181
forerunner cell group	0.374
blastoderm	0.343
presumptive blood	0.350
presumptive mesoderm	0.381
yolk	0.480
anatomical structure	0.818
axis	0.927

The most significantly enriched GO categories in g:profiler ranged from broad groupings such as embryo development (7.742×10^{-3}), regionalisation (3.047×10^{-3}) and pattern specification processes (5.402×10^{-3}) to anterior/posterior pattern specification (1.029×10^{-5}), somite development (7.104×10^{-5}) and segmentation (7.606×10^{-4}). The latter terms clearly reflect the presence of mesodermal genes in the datasets (*dlc, sp5l, tbxta*) (Figure 5.38).



id	source	term id	term name	p_{adj} (query_1)
9	GO:BP	GO:0009952	anterior/posterior pattern specification	1.029×10^{-5}
12	GO:BP	GO:0061053	somite development	7.104×10^{-5}
7	GO:BP	GO:0003002	regionalization	3.047×10^{-4}
8	GO:BP	GO:0001756	somitogenesis	5.278×10^{-4}
10	GO:BP	GO:0035282	segmentation	7.606×10^{-4}
11	GO:BP	GO:0036342	post-anal tail morphogenesis	2.692×10^{-3}
4	GO:BP	GO:0007389	pattern specification process	5.402×10^{-3}
2	GO:BP	GO:0007498	mesoderm development	6.489×10^{-3}
3	GO:BP	GO:0009790	embryo development	7.742×10^{-3}
13	GO:BP	GO:0060429	epithelium development	1.038×10^{-2}
6	GO:BP	GO:0009888	tissue development	1.192×10^{-2}
1	GO:BP	GO:0001707	mesoderm formation	1.560×10^{-2}
15	GO:BP	GO:0048793	pronephros development	2.205×10^{-2}
14	GO:BP	GO:0048332	mesoderm morphogenesis	2.272×10^{-2}
5	GO:BP	GO:0007275	multicellular organism development	4.085×10^{-2}

Figure 5.38 Manhattan plot for significantly downregulated genes in *sox32*$^{-/-}$. Node size is proportional to the total number of genes within each gene set. Strong enrichment from derivatives of mesodermal genes was observed.

Overall, the results from this enrichment analysis confirmed that the downregulated genes following Sox32 mutation were also involved in mesodermal fate. It appears that at the beginning of gastrulation, Sox32 is implicated in specifying which cells express particular genes and regulating which specific mesodermal fate the cells should acquire (*bmp4*, *her1*, *eve1*, *pcdh8*, *dlc*, *dld*, *notch*, *her7*, *tbxta*, *sp5l*). In addition, Sox32 also regulates endodermal genes implicated in cell migration (*cdh6*, *pcdh8*, *mcelsr1a*), Nodal signalling (*mxtx2*, *nanog*, *dusp6*), forerunner cells (*nrip1a*, *rasgef1ba*, *pkd2*, *sp5l*, *fibpb*, *ccdc103*, *mns1*) and future endodermal structures (*sox17*, *aldh1a2*, *lcp1*, *hkdc1*, *cdx4*). The downregulation of some of these genes was already reported in previous publications (Aoki et al., 2002; Dickmeis et al., 2001; Kikuchi et al., 2001; Kinkel et al., 2008; Perez-Camps et al., 2016; Poulain et al., 2006; Xu et al., 2012; Yamamoto et al., 1998); nevertheless, I have obtained more detailed results for the enriched terms and identified new candidate genes downstream of Sox32 at this time point during gastrulation.

5.6.4 Differential expression analysis 9.00 hpf

Differential expression analysis at 9.00 hpf revealed 297 DEGs ($log_2FC > 1$ and FDR < 0.05) between the *sox32*$^{-/-}$ and WT condition, including 108 upregulated and 189 downregulated genes. The top downregulated genes included known endodermal genes such *sox17*, *gata5* and *foxa2*. Notably, *cdh6* was downregulated in the mutant again, at 9.00 hpf. *mixl1* and *cxcl12b* were among the upregulated genes. A volcano plot for the differentially expressed genes is shown in Figure 5.39.

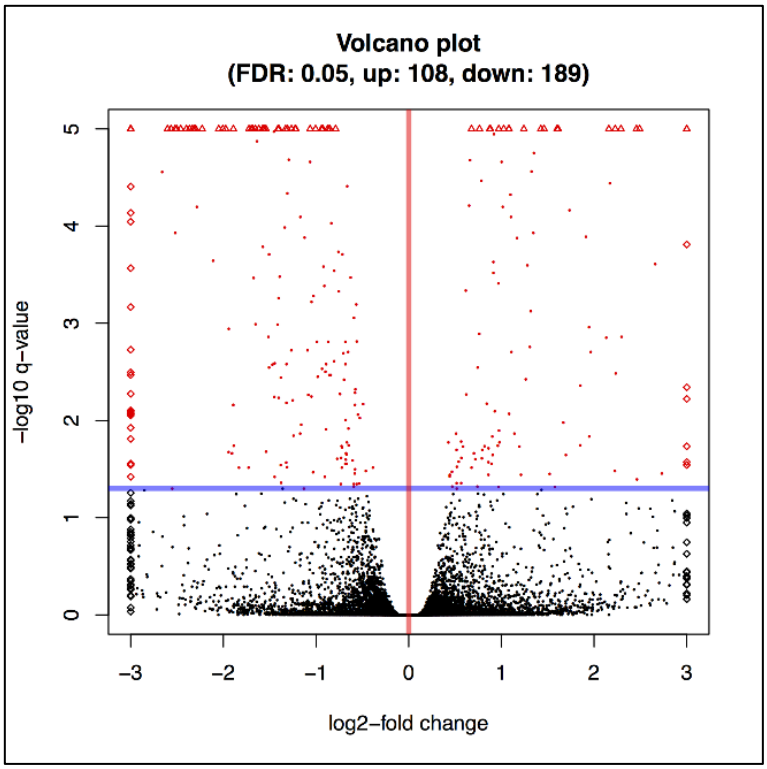


Figure 5.39 Volcano plot of DEGs of *sox32*$^{-/-}$ at 9.00 hpf. The plot displayed the significantly altered genes in red that were identified between the mutant and WT datasets. Blue horizontal line is the FDR threshold. Red vertical line separates positive/negative values.

The relatively low number of DEGs candidates was unsurprising and most likely reflected the higher commitment to endodermal fate at 9.00 hpf than the mesendodermal stage at 5.25 hpf.

A heatmap of the top 150 DEGs revealed a clear distinction between mutant and WT samples (Figure 5.40). Interestingly, this core set of genes also resulted in differential clustering of gene subtypes, suggesting that the core genes may have 2 different expression levels associated with differing roles; one associated to migration of cells and the other associated with endodermal and mesodermal cell identity. In particular not only genes

associated with morphogenetic movements of gastrulation and linked to actin mobility and adhesion modulation were differentially expressed in *sox32*$^{-/-}$ but also genes related to pharyngeal arch cell lineage (*txn*, *met*, *ednraa*, *vwf*, *flrt3*), hearth primordium (*gata5*, *casz1*), pancreas primordium (*foxa2*, *jag1a*) and pronephric primordium (*mnx2b, cdh6*, *prdx5*, *met*, *foxj1a*, *ahi1*, *acsl1b*, *grhl2b*, *ahcyl2*, *dnah9*, *spag6*). Thus, Sox32 was important in patterning the specification and migration of progenitors of multiple different mesodermal and endodermal structures.

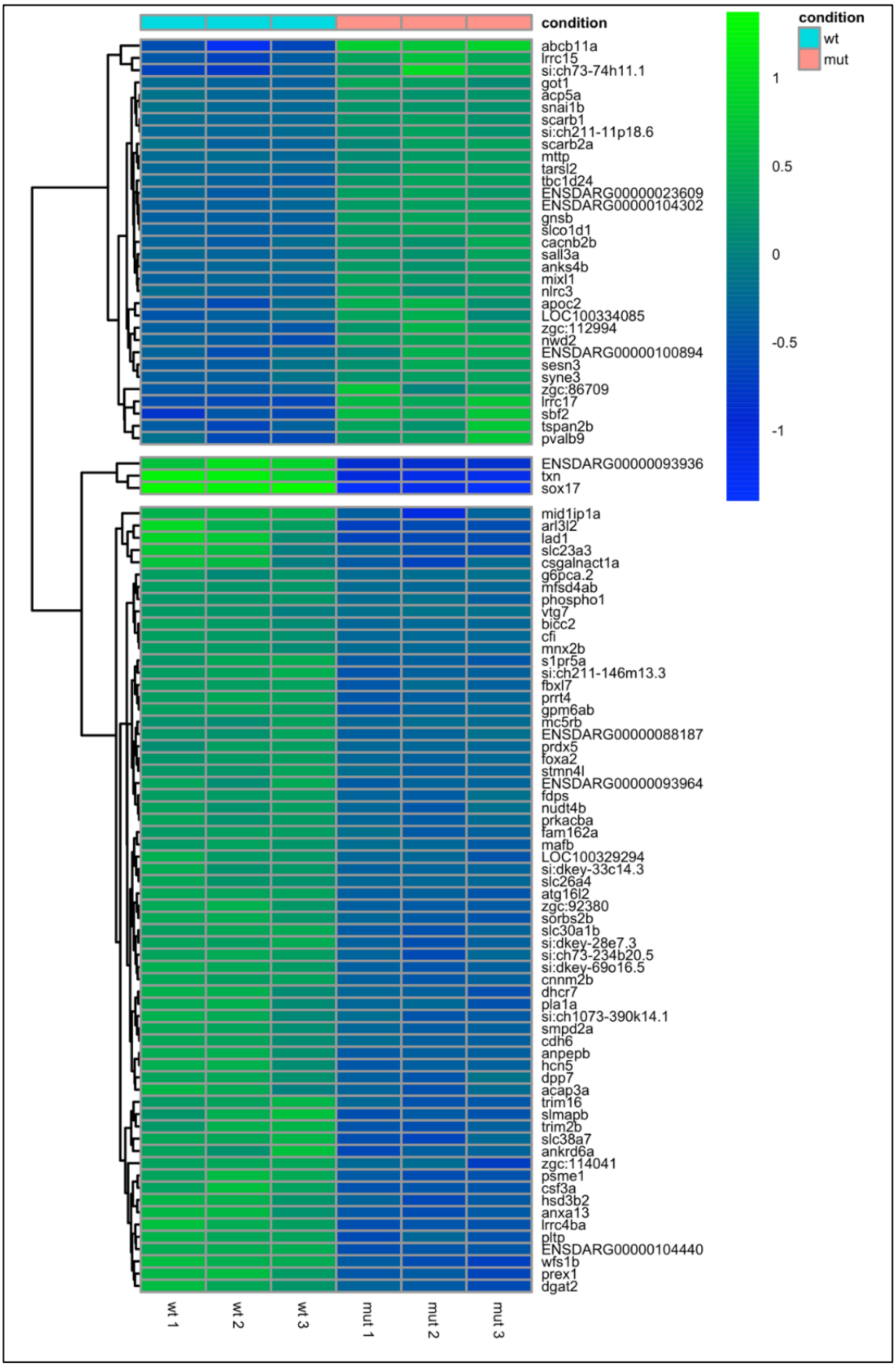


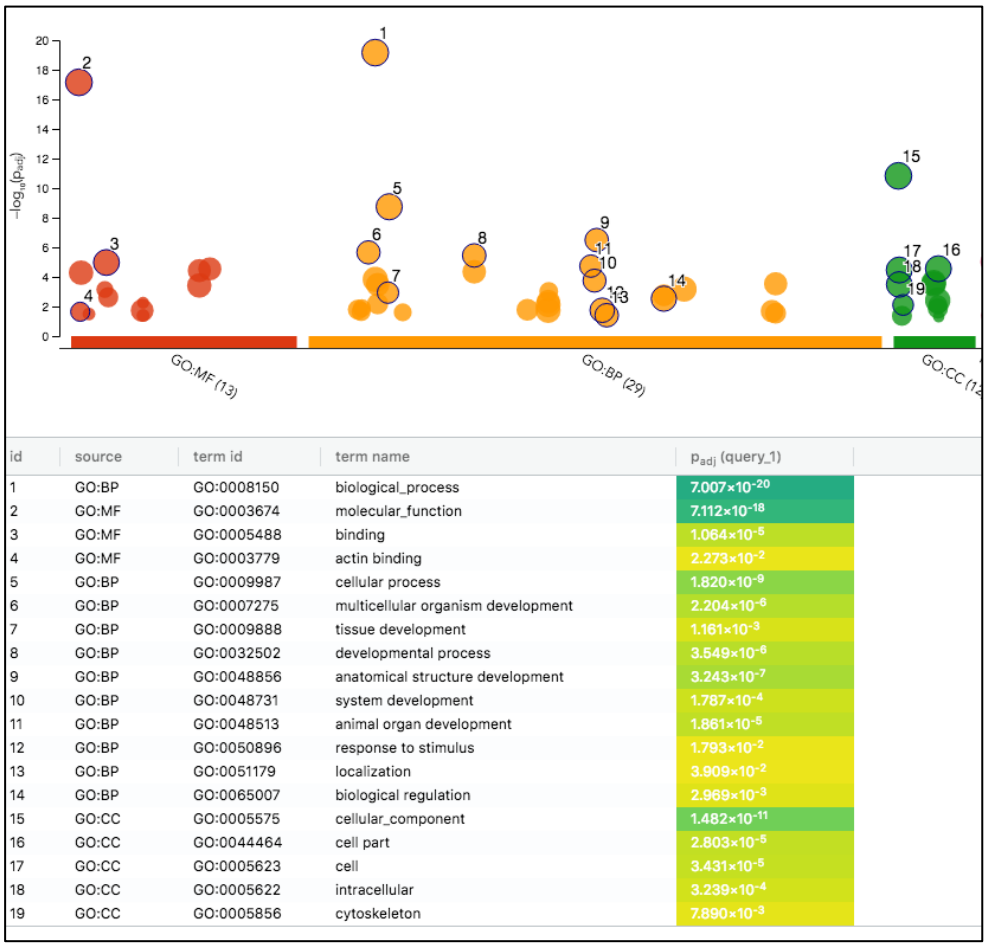
Figure 5.40 Heatmap summarising the top 150 DEGs in *sox32*$^{-/-}$ at 9.00 hpf. Among the most

upregulated gene in the mutant was *mixl1*. Among the downregulated genes a clear endodermal expression signature was observable with *gata5*, *foxa2*, *sox17*, *prex1*, *cxcr4a* and *mnx2b*.

5.6.5 Enrichment analysis at 9.00 hpf

To elucidate the biological roles of the DEGs I then proceeded to functional enrichment analysis. I used the list of downregulated genes in the *sox32*$^{-/-}$ embryos obtained from the RNA-seq analysis to explore the ZFIN database looking for information available on their expression patterns. As expected, spatial information on gene expression from ZEOGS results matched the existing knowledge in the literature, with *sox32*$^{-/-}$ downregulated genes being associated with endodermal cells ($p < 0.01$), endoderm ($p < 0.05$), forerunner cells ($p < 0.05$) and mesodermal fates ($p < 0.01$).

The results from g:profiler associated the genes to broad categories including, molecular function (7.112×10^{-18}), biological process (7.007×10^{-20}), binding (1.064×10^{-5}), cellular process (1.820×10^{-9}), multicellular organism development (2.204×10^{-6}), developmental process (3.549×10^{-6}), anatomical structure development (3.243×10^{-7}), system development (1.787×10^{-4}), animal organ development (1.861×10^{-5}), cellular component (1.482×10^{-11}) and cytoskeleton (7.890×10^{-3}) (Figure 5.41).



id	source	term id	term name	p_{adj} (query_1)
1	GO:BP	GO:0008150	biological_process	7.007×10^{-20}
2	GO:MF	GO:0003674	molecular_function	7.112×10^{-18}
3	GO:MF	GO:0005488	binding	1.064×10^{-5}
4	GO:MF	GO:0003779	actin binding	2.273×10^{-2}
5	GO:BP	GO:0009987	cellular process	1.820×10^{-9}
6	GO:BP	GO:0007275	multicellular organism development	2.204×10^{-6}
7	GO:BP	GO:0009888	tissue development	1.161×10^{-3}
8	GO:BP	GO:0032502	developmental process	3.549×10^{-6}
9	GO:BP	GO:0048856	anatomical structure development	3.243×10^{-7}
10	GO:BP	GO:0048731	system development	1.787×10^{-4}
11	GO:BP	GO:0048513	animal organ development	1.861×10^{-5}
12	GO:BP	GO:0050896	response to stimulus	1.793×10^{-2}
13	GO:BP	GO:0051179	localization	3.909×10^{-2}
14	GO:BP	GO:0065007	biological regulation	2.969×10^{-3}
15	GO:CC	GO:0005575	cellular_component	1.482×10^{-11}
16	GO:CC	GO:0044464	cell part	2.803×10^{-5}
17	GO:CC	GO:0005623	cell	3.431×10^{-5}
18	GO:CC	GO:0005622	intracellular	3.239×10^{-4}
19	GO:CC	GO:0005856	cytoskeleton	7.890×10^{-3}

Figure 5.41 Manhattan plot for significantly downregulated genes in *sox32*$^{-/-}$. The analysis outlined broad functions associated with the top 100 downregulated genes.

The next step was to validate the predictions from the RNA-seq analysis, hence I collected 3 additional biological replicates of *sox32*$^{-/-}$ embryos and WT siblings at both 5.25 hpf and 9.00 hpf to confirm the expression patterns of predicted DEGs by RT-qPCR. I decided to validate 11 genes at 5.25 hpf with 2 (*mxtx2* and *nanog*) upregulated in the RNA-seq datasets and 9 downregulated (*cdh6*, *cdx4*, *dusp4*, *dusp6*, *dlc*, *tbxta*, *sp5l*, *her1*, *eve1*). I also validated 20 genes at 9.00 hpf, all downregulated. The expression values were all normalised to the housekeeping gene *elf2* and expressed as fold change to expression in WT.

As shown in Figure 5.42, 8 out of 11 genes were significantly differentially expressed in the 2 conditions as predicted by the RNA-seq analysis at 5.25 hpf. In contrast, 3 genes (*dusp4, sp5l, eve1*) that had significantly statistically expression in WT than in mutant in the RNA-seq analysis did not show the same significance in the RT-qPCR data.

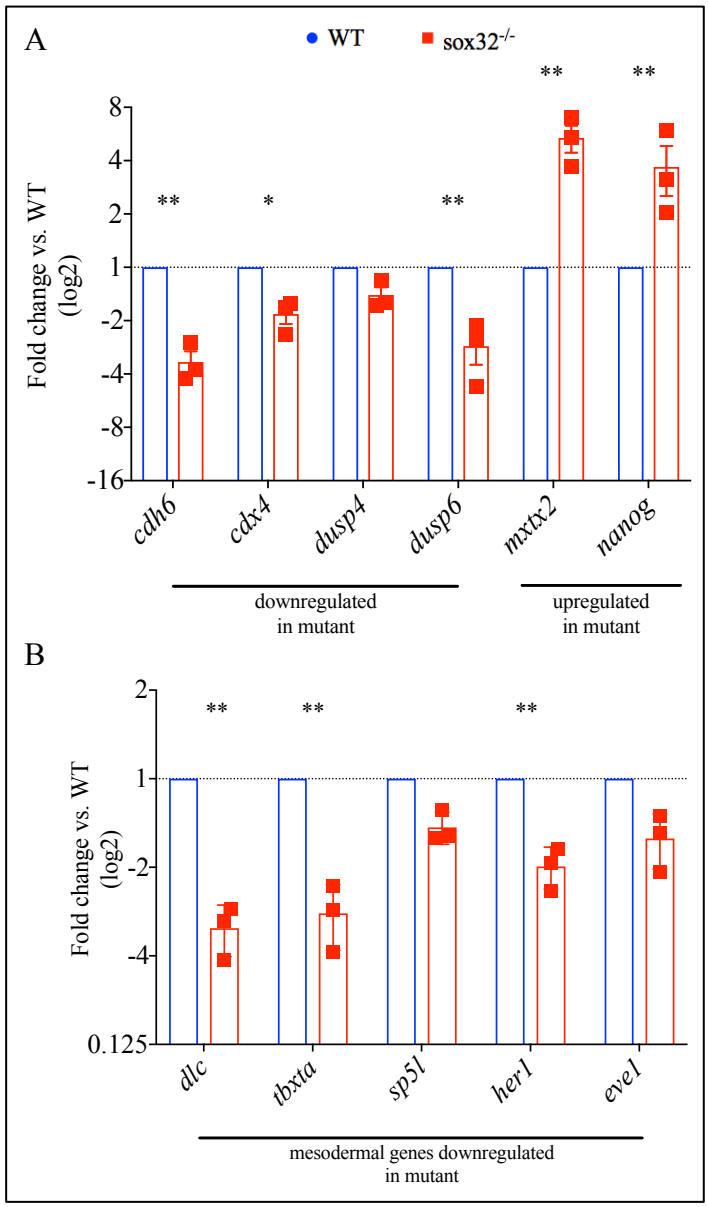


Figure 5.42 RT-qPCR validation of differentially expressed genes in *sox32*$^{-/-}$ at 5.25 hpf.
Comparison of the expression of 11 genes in *sox32*$^{-/-}$ (red bars) and WT (blue bars) using RT-qPCR. (**A**) Endodermal genes and (**B**) mesodermal genes. Fold differences in expression as compared to WT. Positive values indicate higher gene expression in the mutant while negative values indicate higher expression values in the WT. Fold differences were all calculated in $\log_2$. Unpaired two-tailed t-test *$p \leq 0.05$, **$p \leq 0.01$ (n= 3).

From these results, I concluded that at the beginning of gastrulation, Sox32 plays a role in regulating Nodal signalling which diffuses from the YSL acting on both *nanog* and *mxtx2*. Sox32 possibly also regulates the Nodal/Fgf interplay through Dusp4 and Dusp6. In addition, Sox32 regulates, either directly or indirectly, the expression of multiple mesodermal genes (*dlc*, *tbxta*, *sp5l*, *her1*, *eve1*) and simultaneously starts to coordinate the expression of cell-cell adhesion molecules proteins (*cdh6*, *cdx4*) which are associated with actin bundles at the

cytoplasmic domain and intrinsically involved in contractility and, consequently, endodermal cell movement. It would be interesting to investigate the role of additional contractility proteins in these mutants in future work.

As shown in Figure 5.43, most of the DEGs predicted by the RNA-seq analysis were significantly downregulated in the sox32$^{-/-}$. At 9.00 hpf, the effect of the mutation was more severe at the molecular (expression) level than the previous time point as shown by the higher fold changes in the RT-qPCR data.

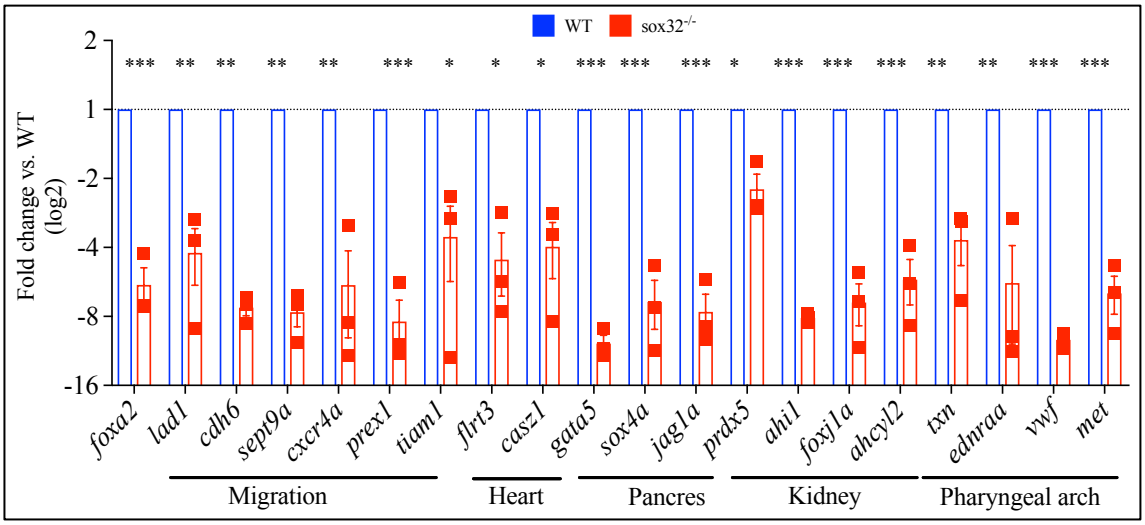


Figure 5.43 RT-qPCR validation of differential expressed genes 9.00 hpf. Comparison of the expression of 20 genes in *sox32*$^{-/-}$ and WT using RT-qPCR. Fold differences in expression as compared to WT. Genes were grouped on the functional activity: migration, heart formation, liver formation, kidney formation and pharyngeal arc development. Fold differences are all calculated in $\log_2$. Unpaired two-tailed t-test *$p \leq 0.05$, **$p \leq 0.01$(n= 3).

The results summarised the genome-wide average pleiotropic effects of Sox32 which operate through independent biological pathways specifying multiple cell fates: promoting endodermal cell motility through the activity of *lad1*, *cdh6*, *sept9a* and *prex1*, coordinating heart specification with *casz1* and *gata5*, and regulating both liver, kidney and pharyngeal arch derivates. Hence, Sox32 is an important TFs in all endodermal structures in zebrafish.

5.6.6 Comparison of 5.25 and 9.00 hpf *sox32*$^{-/-}$ transcriptome

My current study had looked at 2 stages of development, at the beginning of gastrulation and at the end, the latter being just prior to the start of somitogenesis and the expression of tissue specific genes. I next therefore asked how similar the transcriptomes were at the 2 time points to enable me to outline similarities and differences in Sox32 downstream effectors at the different time points. Table 5.10 and the associated Venn diagram (Figure 5.44) depict the total number of genes identified in the corresponding upregulated and downregulated datasets.

Table 5.10 Number of DEGs in 5.25 and 9.00 hpf *sox32*$^{-/-}$ transcriptome with $log_2FC > 1$ and FDR < 0.05.

Developmental time	Contrast	Total No. of Significant DEGs	No. of Upregulated Genes	No. of Downregulated Genes
5.25 hpf	Sox32$^{-/-}$ vs. Wt (with batch effect)	8	0	8
5.25 hpf	Sox32$^{-/-}$ vs. Wt (without batch effect)	444	192	253
9.00 hpf	Sox32$^{-/-}$ vs. Wt	297	108	189

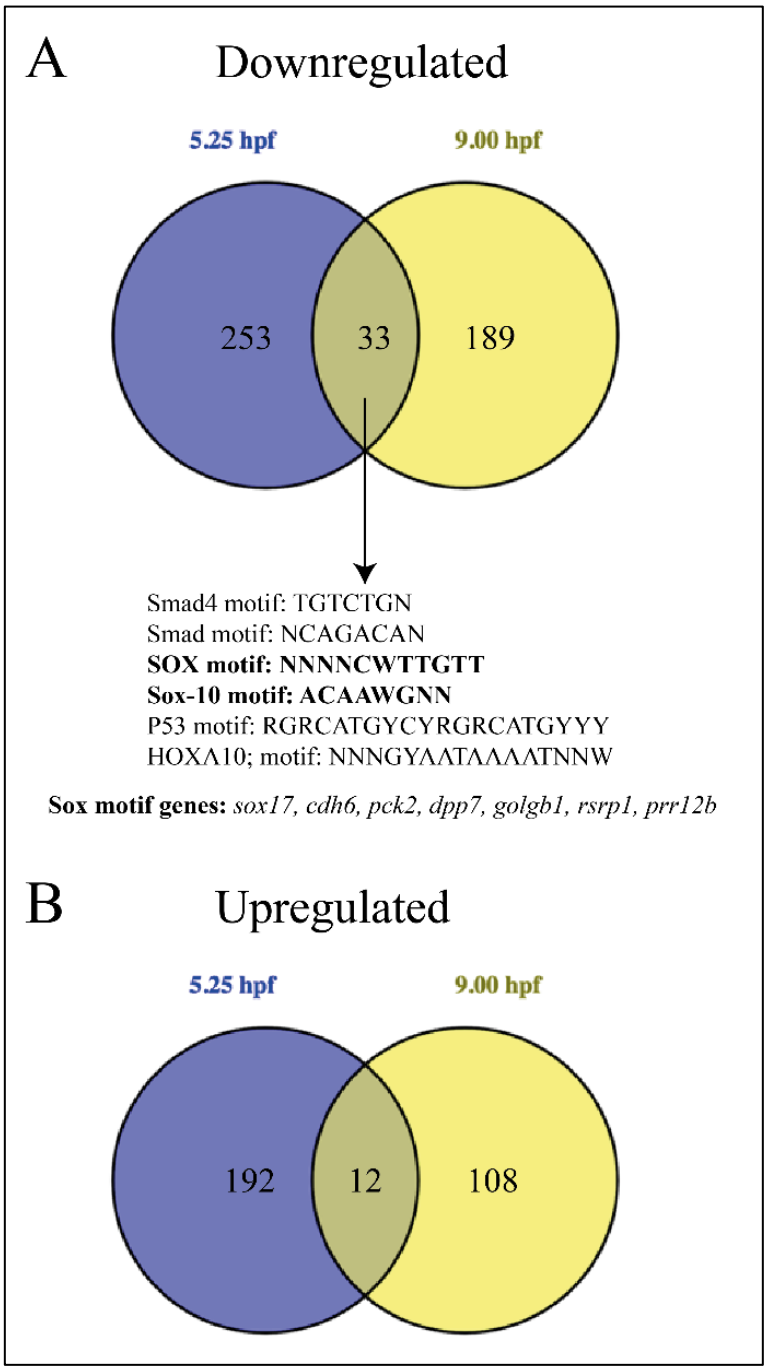


Figure 5.44 Common genes in the *sox32*$^{-/-}$ transcriptomes at 5.25 (after batch correction) and 9.00 hpf. Venn diagram showing the number of core DEGs, separated by (**A**) downregulation and (**B**) upregulation. The common 33 downregulated genes were characterised by sharing a Smad, P53 and Sox motif.

Position weight matrix (PWM) containing putative transcription factor binding sites were download from TRANSFAC database (Matys et al., 2006) and input in g:profiler to search for motifs 1000 bp upstream of TSS of query genes. No common motif was identified in the upregulated overlapping genes, but reassuringly, the 2 time points shared 33 common downregulated genes which were associated by g:profiler to 3 distinct motifs: Smad, Sox and

P53. Smad proteins are responsible for transducing Nodal signals into the nucleus and therefore these results reinforce the critical role played during gastrulation by Nodal signalling which drives cells towards mesendodermal fate. The Sox motif was interesting, particularly because I could speculate that most of these 33 genes were directly bound by Sox32 and their differential expression was not an indirect effect of Sox32 binding another TFs. Lastly, P53 has been linked to programmed cell death and apoptosis (Nikolay Popgeorgiev, 2018), side effects that could be caused by the *sox32* mutation. A decrease in cell survival at the end of gastrulation has also been observed in MZnanog mutant zebrafish embryos (Veil et al., 2018). Hence, Sox32 is an important TF that coordinates Nodal signalling and plays a key role in the proliferation and survival of endodermal cells.

5.7 *sox17:GFP* transcriptome

One of the limitations of the RNA-seq datasets just described was that they encompassed the transcriptome of the whole embryo and not just the endodermal cells. I therefore used the endoderm specific transgenic line described in Chapter 4, where GFP is under the control of the *sox17* promoter, to sort GFP positive cells by flow cytometry at 9.00 hpf, the end of gastrulation. I then took a snapshot of the transcriptome of endoderm cells by RNA-seq, allowing me to obtain a well-defined transcriptional signature representing the developing endoderm.

5.7.1 RNA isolation optimization for *sox17:GFP* cells

I have already described the process used to isolate and extract high quality RNA from sorted cells, which consist of a relatively small population of cells in the developing zebrafish (~25%) at 9.00 hpf (see Chapter 4). I now further describe the optimization process I used to ensure high cell viability after dissociation and suspension of cells, followed by high yield isolation of intact RNA.

My protocol yielded an average of 60.1% viable cells (n=12) ranging from 51.8% to 69.0%. On average, I collected 50,000 GFP^+ cells per batch for RT-qPCR assessment and 100,000 cells for RNA-seq library preparation. Isolated RNA was analysed for quality and the concentration determined for RNA-seq using Qubit and Agilent 2,100 BioAnalyzer. The RNA concentration for each sample ranged from 140 to 320 ng/batch of sorted embryos (n=8). The protocol generated RNA with RIN values ranging from 7.0 to 8.9, with an average RIN of 8.1

(n=12). RIN values were higher when sorting time was faster (< 60 mins) and RNA was extracted by a combination of TRIzol + BCP (see Chapter 2 Materials and Methods). In my hands, the use of 2 commercially available columns used according to the manufacturers' protocols yielded less total RNA (Direct-zol RNA Microprep and RNeasy Micro Kit) and lower quality RNA than TRIzol and BCP. More specifically, TRIzol outperformed the Qiagen RTL buffer for 50,000 cells both in yield and in the quality of extracted RNA, as shown in Figure 5.45, where sharper peaks were visible for the TRIzol protocol, indicating that the RNA was less degraded. Both protocols needed DNAse treatment to eliminate gDNA contamination.

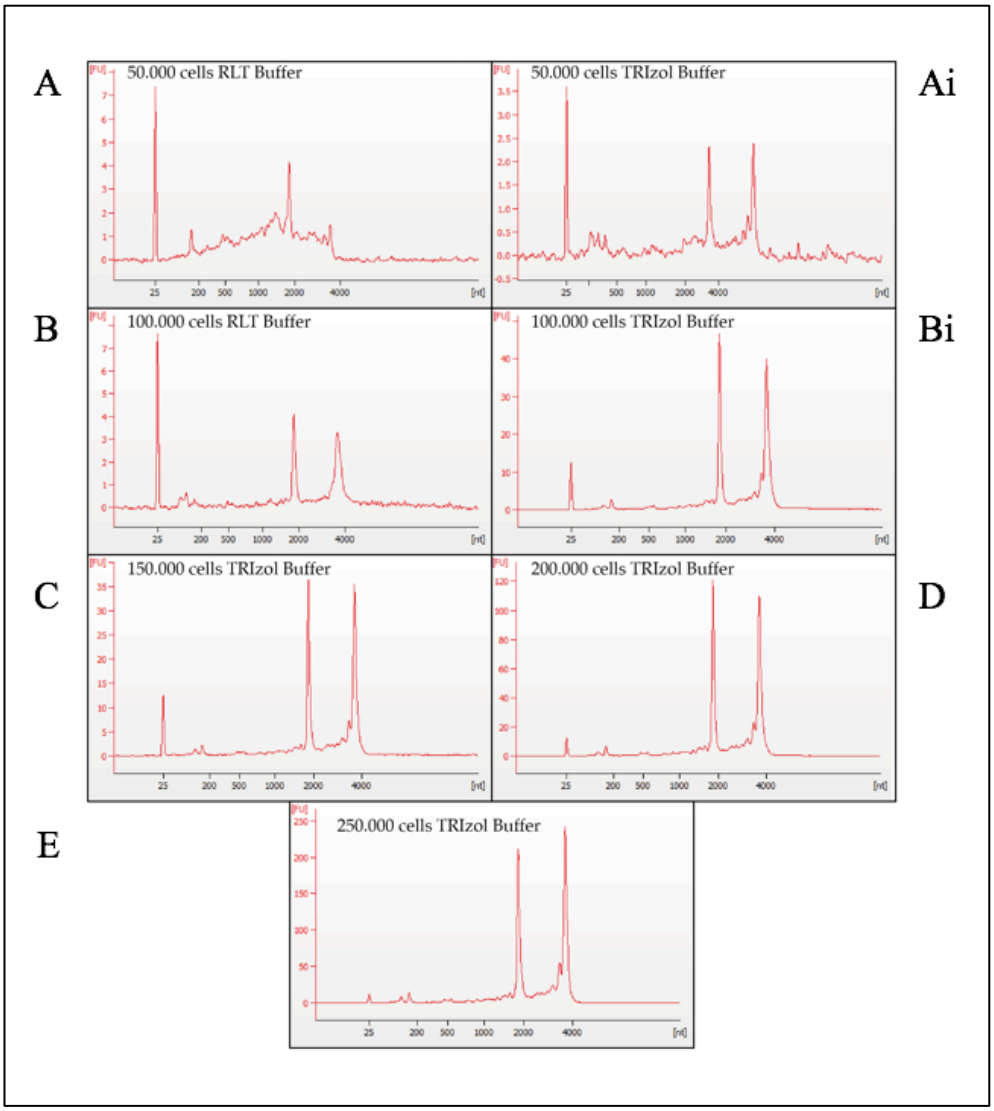


Figure 5.45 RNA quality control. Examples of Bioanalyzer reports of total RNA extracted from different FAC sorted samples using different protocols. (**A**) 50,000 cells extracted using RLT buffer for lysing the cells prior to RNA isolation with Qiagen kit and (**Ai**) TRIzol protocol. (**B**) Use of Qiagen kit and (**Bi**) TRIzol to extract RNA from 100,000 cells. Panels (**C**), (**D**) and (**E**) show representative examples of the quality results of

isolated RNA from 150,000, 200,000 and 250,000 cells respectively. Samples were considered to be of high quality if the ribosomal peaks had minimal (< 10% of the total peak area) shoulders as shown starting from 100,000 cells in TRIzol (Bi). All of the samples selected for RNA sequencing met these criteria.

As described in the previous chapter, methods were adopted during the cell sorting process to reduce the risk of introducing artefacts. Specifically, a stringent gating strategy was used, as well as wildtype controls for autofluorescence and DAPI staining for dead/alive cells. Lastly, I performed RT-qPCR on all samples before preparing the RNA-seq libraries to assess the enrichment of *sox32* and *sox17* genes in the GFP$^+$ cell population (Figure 5.46).

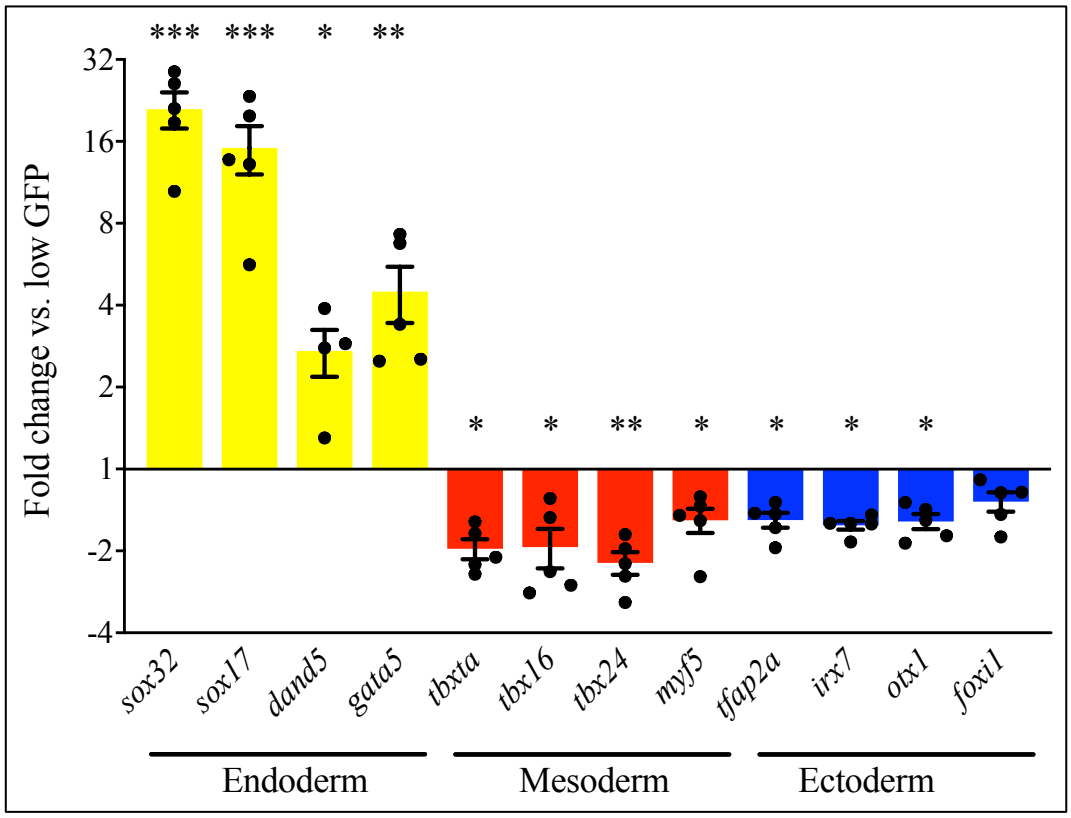


Figure 5.46 RT-qPCR results from FAC sorted *sox17*:*GFP* RNA-seq libraries. RT-qPCR on 5 libraries was preemptively done to check for enrichment of endodermal genes in the GFP$^+$ population, and to test for the degree of contamination of unwanted GFP$^+$ non endodermal cells. The GFP$^+$ sample showed enrichment for the endodermal markers *sox32*, *sox17* and *gata5* whereas *dand5*, a marker for forerunner cells, was detected at lower levels. Mesodermal and ectodermal marker expression was detected at background level in GFP$^-$ cells. Unpaired t-test *$p \leq 0.05$, **$p \leq 0.01$, ***$p \leq 0.001$. Mean ± SEM (n=5).

A summary of library characteristics is provided in Table 5.11. Read depth was comparable between GFP$^-$ and GFP$^+$ samples.

Table 5.11 Summary of total reads for *sox17:GFP* libraries.

	GFP$^-$					
	Rep 1	Rep 2	Rep 3	Rep 4	Rep 5	Rep 6
Total reads	11.6 M	18.2 M	16.1 M	12.6 M	17.3 M	19.8 M
	GFP$^+$					
	Rep 1	Rep 2	Rep 3	Rep 4	Rep 5	Rep 6
Total reads	16.3 M	17.8 M	19.4 M	17.8 M	17.4 M	14.2 M

5.7.2 A large set of genes (>10%) were specifically differentially expressed in endodermal cells.

Using a cutoff of absolute FC > 1 and an FDR < 0.05, I found that 442 genes were significantly differentially expressed in GFP$^+$ compared to GFP$^-$ cells. The PCA and clustering plots showed the relationships between the replicates (Figure 5.47). Both analyses demonstrated how well replicates segregate by condition, outlining the reproducibility and similarity in gene expression of the replicates which was high considering the numerous steps required to obtain libraries from FAC sorted embryos. Nonetheless, some variation between samples (for example samples GFP$^-$1 showed lower reproducibility than others) indicated that a degree of variability existed (e.g. duration of sorting and collecting the cells, slightly different staging of the embryos) and that my choice to have more than 3 replicates was rational and sensible to achieve good statistical results. As shown by the PCA plot, 4 biological replicates of GFP$^+$ cells were reproducible as they were all clustered together with small variability. The 3 data points of the GFP$^-$ cells were also highly reproducible (Figure 5.47). The large 'distance' that separated the GFP$^+$ and GFP$^-$ cells suggested that their gene expression profiles were quite different, as confirmed by differential gene expression.

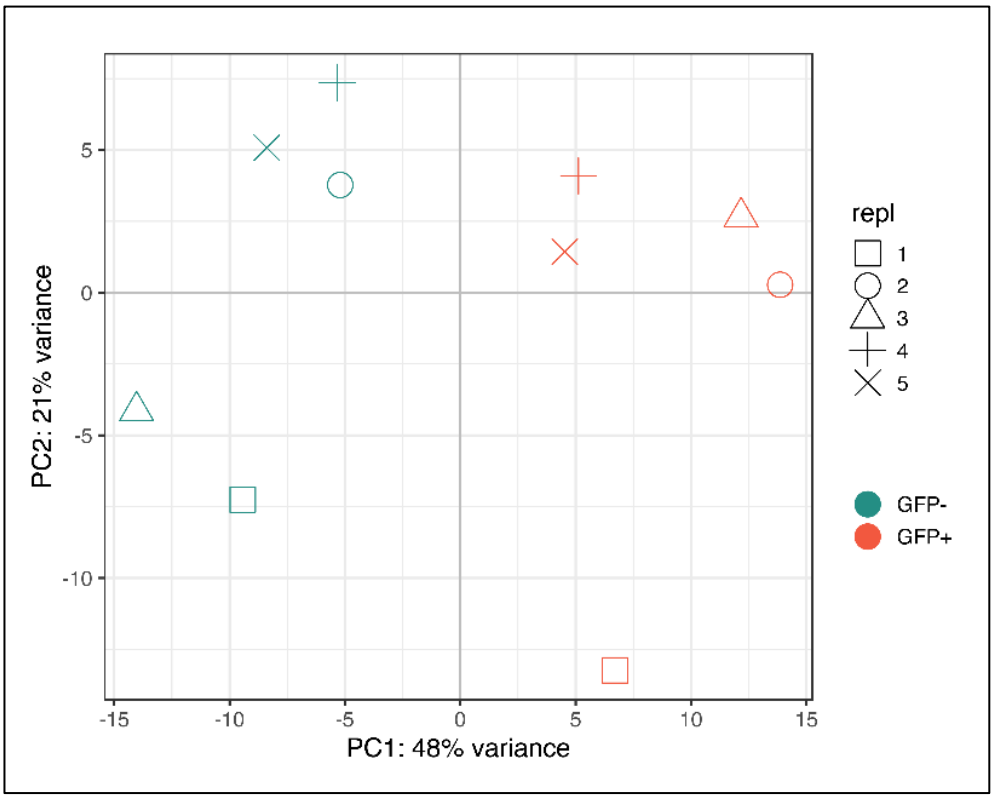


Figure 5.47 Reproducibility between *sox17:GFP* replicates. Principal component analysis (PCA) showing the relationships between the samples and summarising the variation between GFP^+ (red symbols) and GFP^- replicates (green symbols). PCA separated samples from the different biological conditions, denoting that the biological variability (GFP^+ - enriched endodermal population) was the main source of variance in the data (48%).

The DEGs and heatmap analyses showed a clear difference in gene expression between GFP^+ and GFP^- cells, further confirming the role of *sox17* as a marker of endodermal signature and regulator/mediator of endoderm development. As expected, the majority of these genes (349 genes, 79%) had higher levels of expression in GFP^+ than in GFP^- cells with a maximum of +5-fold decrease; a small number of these genes (93 genes, 11%) had higher levels of expression in GFP^- cells compared to GFP^+ cells (Figure 5.48). The number of DEGs was similar to what has been recently reported by Yuan et al. (2018) when cells labelled by the GFP:*Smarcd3*-F6 enhancer, an early marker of cardiac lineages, were sorted at 10.00 hpf which revealed 316 genes differentially expressed between the GFP^+ and GFP^- cell populations.

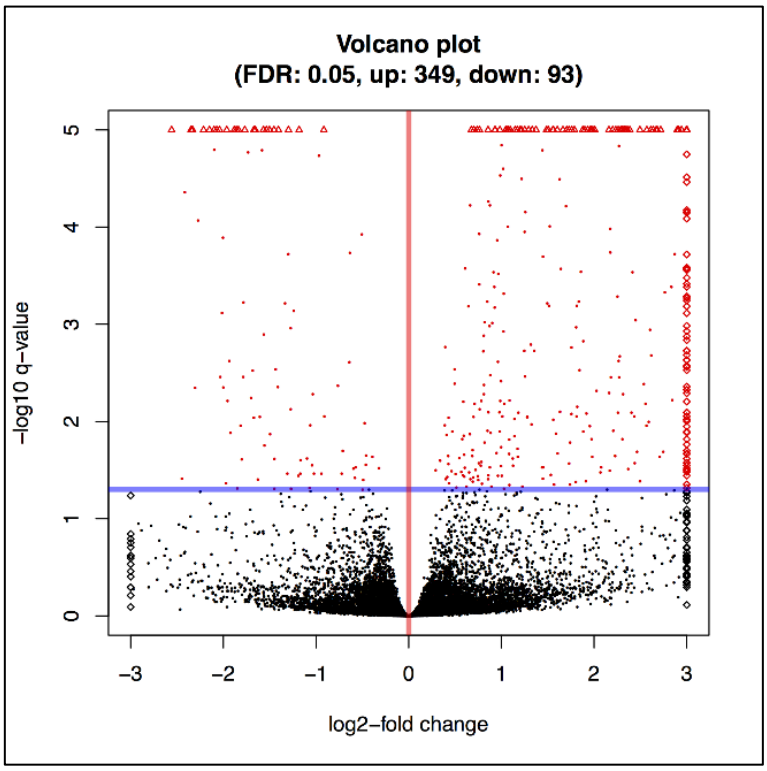


Figure 5.48 *sox17:GFP* volcano plot. Volcano plot showing the log_2FC on the x-axis and the log_{10} FDR values on the y-axis. Genes with significant difference are depicted in red, black indicates genes that did not show significant differences in expression between GFP^+ vs GFP^-. Most of the DEGs were upregulated in GFP^+ cells, positive FC were associated with endodermal signature. The blue horizontal line shows FDR cutoff of value of 0.05. Triangles represents genes whose fold change was too high to be plotted.

Unsupervised clustering using the top 150 genes differentially expressed between GFP^+ and GFP^- populations grouped cells into 3 broad clusters (Figure 5.49). One cluster represented ectodermal (*krt5*$^+$, *krt17*$^+$, *sox2*$^+$) populations highly enriched in the GFP^- cells and 2 clusters of putative endodermal and mesodermal populations highly enriched in the GFP^+ cells. Within this cluster of upregulated genes in GFP^+ cells, I recognised 2 potential subgroups, one coexpressing the known endodermal genes (*sox32*$^+$, *sox17*$^+$, *mnx2b*$^+$, *prex1*$^+$) and one coexpressing both known endodermal, forerunner cell and mesodermal genes (*gata5*$^+$, *eml2*$^+$, *itga5*$^+$, *fgfrl1b*$^+$, *flrt3*$^+$, *tbx1*$^+$, *sept9a*$^+$, *spag6*$^+$, *daw1*$^+$). I also observed the prevalence of genes associated with morphogenetic movement, in particular, the GFP^+ cells were enriched for N-cadherin and Rho GTPase genes (*cdh6*$^+$, *tiam1*$^+$, *prex1*$^+$, *cdh12a*$^+$, *dnah6*$^+$, *ctnnd2b*$^+$, *apln*$^+$, *arhgap36*$^+$) which have been previously linked to control endodermal cell motility and regulate the process of convergence of endoderm and organ precursors toward the embryonic midline in the zebrafish embryo (Babb and Marrs, 2004; Giger and David, 2017; Straub et al., 2011; Warga and Kane, 2007; Woo et al., 2012). My results add

information on the molecular mechanisms that regulate this process, which, however, remain largely unexplored.

Overall, the transcriptional profiles in the GFP^+ and GFP^- specific cell populations further indicated that the *sox17*:GFP transgenic line marks endodermal progenitor cells, and can be exploited to identify novel genes that are commonly regulated, or biological signatures associated with a endoderm development.

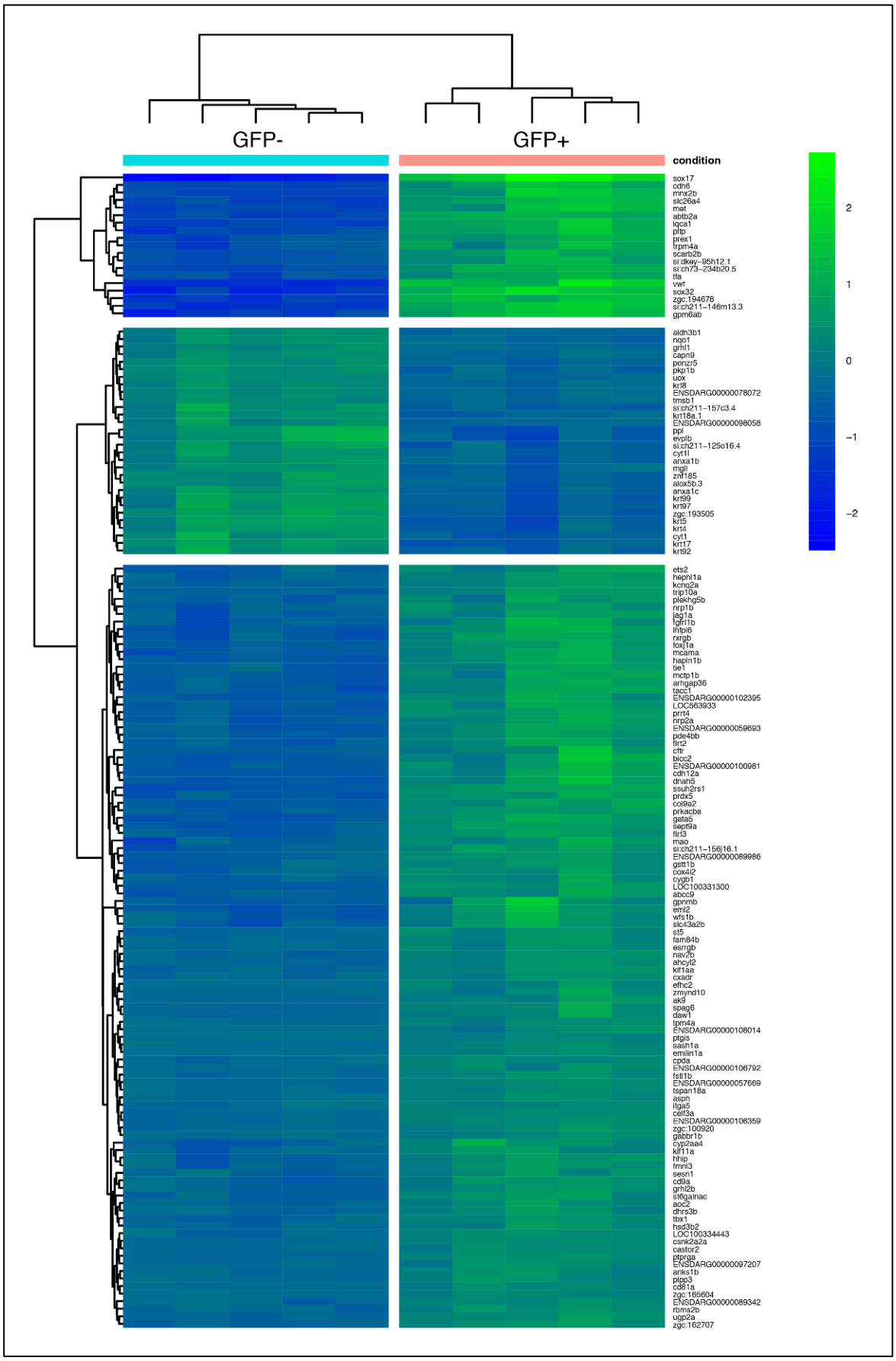
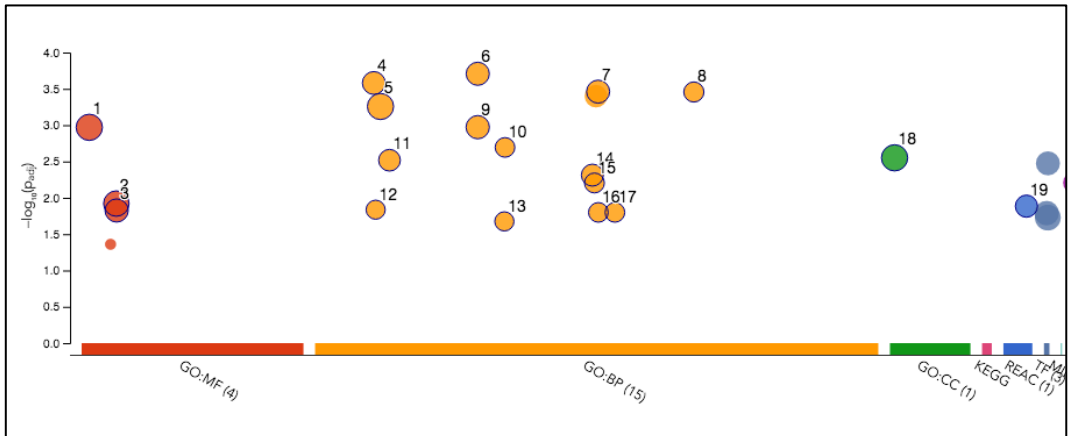


Figure 5.49 Heatmap of the top 150 genes ranked by FDR in *sox17:GFP* RNA-seq. Heatmap and

hierarchical clustering of genes showing statistically significant changes in gene expression. Green indicates high expression and blue indicates low expression. The clustering method which groups genes together based on the similarity of their expression patterns identified significantly higher expression of endoderm markers (*sox32, sox17, foxa2, foxa3*), and known cardiac markers (*gata5, tbx1*) in *sox17:GFP* labelled cells, whilst ectoderm (*krt4*, *krt7*, *gata3*, *tfap2a*) and axial mesoderm (*myod1, grnb, fgf8a*) genes were relatively depleted.

5.7.3 Enrichment analysis showed clear distinction of biological processes and pathways in the endodermal cells.

Consistent with the above results, genes showing higher expression in the GFP^{+} population were enriched (g:profiler) for processes related to endoderm (tube development 2.026×10^{-3}, tube morphogenesis 2.112×10^{-2}), circulatory system development (3.498×10^{-4}) and cell motility (tube development 1.588×10^{-2}) (Figure 5.50), whereas genes enriched in GFP^{-} cells were enriched for those involved in ectoderm and otic placode development (epidermal cell differentiation (2.359×10^{-5}), epidermis development (1.914×10^{-4}), ectodermal placode formation (8.716×10^{-3}) and otic vesicle morphogenesis (2.275×10^{-2}) (data not shown).



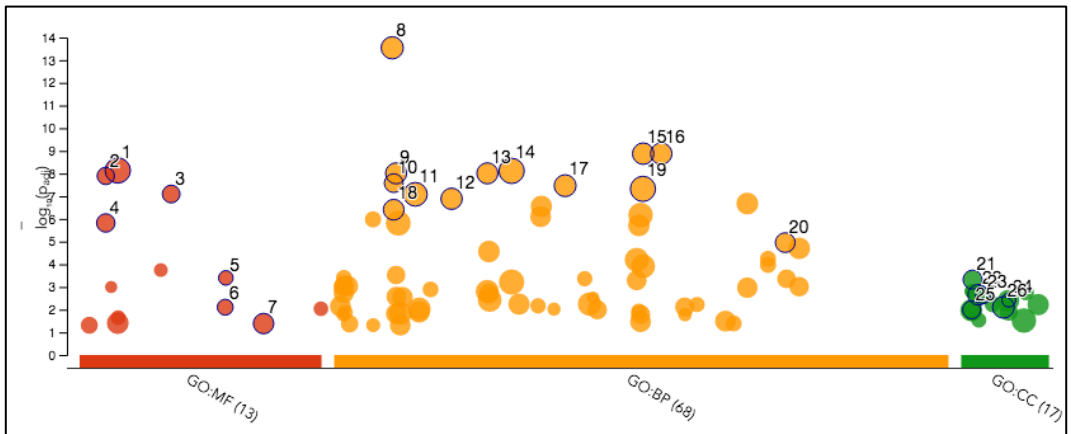
id	source	term id	term name	p_{adj} (query_1)
1	GO:MF	GO:0003674	molecular_function	1.072×10^{-3}
2	GO:MF	GO:0005488	binding	1.201×10^{-2}
3	GO:MF	GO:0005515	protein binding	1.496×10^{-2}
4	GO:BP	GO:0007275	multicellular organism development	2.629×10^{-4}
5	GO:BP	GO:0008150	biological_process	5.535×10^{-4}
6	GO:BP	GO:0032502	developmental process	1.963×10^{-4}
7	GO:BP	GO:0048856	anatomical structure development	3.458×10^{-4}
8	GO:BP	GO:0072359	circulatory system development	3.498×10^{-4}
9	GO:BP	GO:0032501	multicellular organismal process	1.067×10^{-3}
10	GO:BP	GO:0035295	tube development	2.026×10^{-3}
11	GO:BP	GO:0009653	anatomical structure morphogenesis	3.041×10^{-3}
12	GO:BP	GO:0007389	pattern specification process	1.458×10^{-2}
13	GO:BP	GO:0035239	tube morphogenesis	2.112×10^{-2}
14	GO:BP	GO:0048513	animal organ development	4.874×10^{-3}
15	GO:BP	GO:0048646	anatomical structure formation involved in mor...	6.250×10^{-3}
16	GO:BP	GO:0048870	cell motility	1.588×10^{-2}
17	GO:BP	GO:0051674	localization of cell	1.588×10^{-2}
18	GO:CC	GO:0005575	cellular_component	2.818×10^{-3}
19	REAC	REAC:R-DRE-1...	Signal Transduction	1.303×10^{-2}

Figure 5.50 Manhattan plot for significantly upregulated genes in GFP^{+}cells. Node size is proportional

to the total number of genes within each gene set. Molecular function term was enriched for general features such as 'molecular function' and 'protein binding'. Biological processes were enriched for the circulatory system, tube development, morphogenesis and cell mobility. The term associated with the REACTOME dataset was signalling transduction.

Immediately with this analysis it was possible to observe that the top most enriched GO terms obtained from genes that were significantly more highly expressed in the GFP^+ cell populations included circulatory system development (3.498×10^{-4}), tube development (2.026×10^{-3}), tube morphogenesis (2.112×10^{-2}), cell motility (1.588×10^{-2}) and localization of the cell (1.588×10^{-2}). However, observing the gene clustering in Figure 5.49, I was able to separate 2 strong signatures within them: i) genes associated with migration and the cytoskeleton (first cluster) and ii) genes associated with endodermal and mesodermal fate (second cluster), and thus further insights in endodermal cell fate decisions were explicated.

Biological processes enriched in the first cluster not only were related to migration and cytoskeleton (dynein light (3.99×10^{-04}), chain binding ATP-dependent (9.16×10^{-03}), microtubule motor (4.87×10^{-02}), activity microtubule motor activity (4.87×10^{-02}), cell motility (1.33×10^{-09}), microtubule-based movement (2.649×10^{-08}), cell adhesion ($9.79\times^{-}10^{09}$)) but also signalling and cell communication (transmembrane receptor protein tyrosine kinase activity ($1.27\times^{-08}$), cadherin binding ($7.82\times^{-03}$) and GTPase binding ($4.12\times^{-02}$)) which can be associated with the convergence and extension movements of endodermal cells during gastrulation (Figure 5.51) (Babb and Marrs, 2004; Woo et al., 2012).

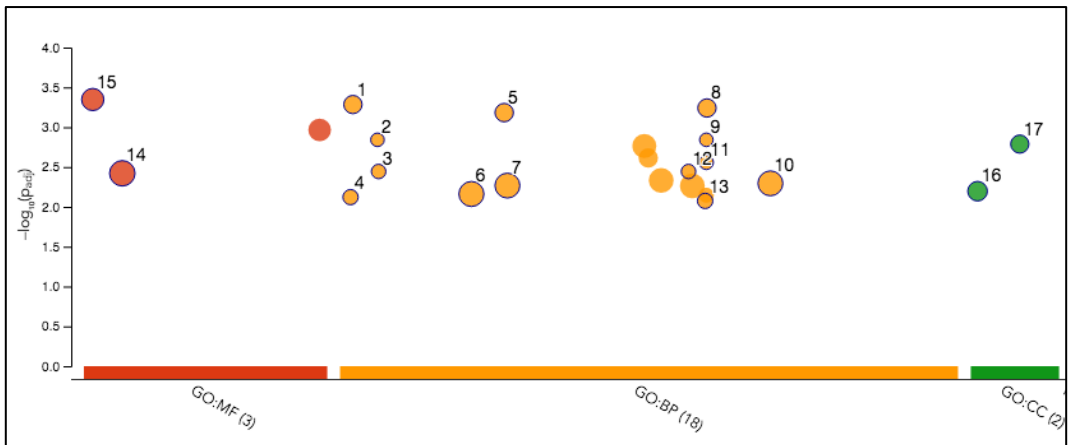


id	source	term id	term name	p_{adj} (query_1)
1	GO:MF	GO:0005515	protein binding	7.302×10^{-9}
2	GO:MF	GO:0004714	transmembrane receptor protein tyrosine kinas...	1.272×10^{-8}
3	GO:MF	GO:0019199	transmembrane receptor protein kinase activity	7.979×10^{-8}
4	GO:MF	GO:0004713	protein tyrosine kinase activity	1.508×10^{-6}
5	GO:MF	GO:0045503	dynein light chain binding	3.993×10^{-4}
6	GO:MF	GO:0045296	cadherin binding	7.818×10^{-3}
7	GO:MF	GO:0051020	GTPase binding	4.115×10^{-2}
8	GO:BP	GO:0006928	movement of cell or subcellular component	2.859×10^{-14}
9	GO:BP	GO:0007155	cell adhesion	9.790×10^{-9}
10	GO:BP	GO:0007018	microtubule-based movement	2.649×10^{-8}
11	GO:BP	GO:0009653	anatomical structure morphogenesis	8.232×10^{-8}
12	GO:BP	GO:0016477	cell migration	1.295×10^{-7}
13	GO:BP	GO:0022610	biological adhesion	9.790×10^{-9}
14	GO:BP	GO:0032502	developmental process	7.640×10^{-9}
15	GO:BP	GO:0048870	cell motility	1.329×10^{-9}
16	GO:BP	GO:0051674	localization of cell	1.329×10^{-9}
17	GO:BP	GO:0040011	locomotion	3.428×10^{-8}
18	GO:BP	GO:0007017	microtubule-based process	3.928×10^{-7}
19	GO:BP	GO:0048856	anatomical structure development	4.742×10^{-8}
20	GO:BP	GO:0098609	cell-cell adhesion	1.111×10^{-5}
21	GO:CC	GO:0005929	cilium	4.756×10^{-4}
22	GO:CC	GO:0015630	microtubule cytoskeleton	2.225×10^{-3}
23	GO:CC	GO:0030286	dynein complex	2.420×10^{-3}
24	GO:CC	GO:0044447	axoneme part	3.790×10^{-3}
25	GO:CC	GO:0005875	microtubule associated complex	1.017×10^{-2}
26	GO:CC	GO:0042995	cell projection	7.971×10^{-3}

Figure 5.51 Manhattan plot for significantly downregulated genes in *GFP*$^{+}$ (first genes cluster). Migration molecular signature was enriched for cytoskeletal regulators involved in contractility and cell-cell adhesion protein. A role for Rho GTPases in single-cell motility and the importance of transmembrane signalling was also highlighted by the enrichment analysis.

Biological processes enriched in the second cluster were mainly related to relationships between endoderm cells, the development of cardiac tissue from mesodermal cells and how endoderm-derived growth factors regulate the formation of both cell fates during specification, and morphogenesis of cells in developing embryos. These processes included: liver development (5.200×10^{-4}), endoderm formation (7.557×10^{-3}), pancreas development 6.567×10^{-4}), hepatobiliary system development (5.744×10^{-4}), heart formation (8.432×10^{-3}). Enriched genes were also connected to protein binding (3.795×10^{-3}), DNA-binding transcription factor activity (4.498×10^{-4}), transcription factor complex (6.363×10^{-3}) and

nuclear transcription factor complex (1.630×10^{-3}) (Figure 5.52). Mesodermal and endodermal cells regulate highly intertwined processes during gastrulation and not surprisingly enrichment for cardioblast migration (1.441×10^{-3}), cell migration to the midline involved in heart development (3.601×10^{-3}), cell migration involved in heart formation (2.773×10^{-3}) and convergent extension involved in organogenesis (3.601×10^{-3}) were also observed in this cluster (David and Rosa, 2001; Sakaguchi et al., 2006).



id	source	term id	term name	p_{adj} (query_1)
1	GO:BP	GO:0001889	liver development	5.200×10^{-4}
2	GO:BP	GO:0003260	cardioblast migration	1.441×10^{-3}
3	GO:BP	GO:0003318	cell migration to the midline involved in heart d...	3.601×10^{-3}
4	GO:BP	GO:0001706	endoderm formation	7.557×10^{-3}
5	GO:BP	GO:0031016	pancreas development	6.567×10^{-4}
6	GO:BP	GO:0019222	regulation of metabolic process	6.900×10^{-3}
7	GO:BP	GO:0031323	regulation of cellular metabolic process	5.413×10^{-3}
8	GO:BP	GO:0061008	hepaticobiliary system development	5.744×10^{-4}
9	GO:BP	GO:0060975	cardioblast migration to the midline involved in...	1.441×10^{-3}
10	GO:BP	GO:0080090	regulation of primary metabolic process	5.073×10^{-3}
11	GO:BP	GO:0060974	cell migration involved in heart formation	2.773×10^{-3}
12	GO:BP	GO:0060029	convergent extension involved in organogenesis	3.601×10^{-3}
13	GO:BP	GO:0060914	heart formation	8.432×10^{-3}
14	GO:MF	GO:0005515	protein binding	3.795×10^{-3}
15	GO:MF	GO:0003700	DNA-binding transcription factor activity	4.498×10^{-4}
16	GO:CC	GO:0005667	transcription factor complex	6.363×10^{-3}
17	GO:CC	GO:0044798	nuclear transcription factor complex	1.630×10^{-3}

Figure 5.52 Manhattan plot for significant upregulated genes in GFP^{+} (second genes cluster). The second cluster was associated to an endodermal and mesodermal molecular signature. Molecular processes were linked to protein binding and DNA-binding transcription factor activity. Cellular processes were related to transcription factor complex and nuclear transcription factor complex. Biological processes were interconnected to both endodermal structures (liver and pancreas formation) and mesodermal structures (cardiac migration and heart formation)

Together, my transcriptomic analyses demonstrated that cells labelled by the *sox17*-GFP promoter were enriched for endodermal and cardiac lineages. These data reinforce the relationship between endodermal cells and the development of cardiac tissue with the putative

progenitor population apparent by as early as the beginning of gastrulation (mesendodermal cells at the margin).

5.7.4 RNA-seq results confirmed by RT-qPCR

To validate the RNA-seq predictions on novel genes in the endodermal progenitors within the 9.00 hpf *sox17*-GFP labelled cells, I selected 17 genes, 15 enriched in GFP^+ population and 2 enriched the GFP^- to test with RT-qPCR analysis. The expression values were all normalised to the cycle threshold (Ct) value of the reference gene *elf2* and plotted as fold change GFP^+/GFP^- (Figure 5.53).

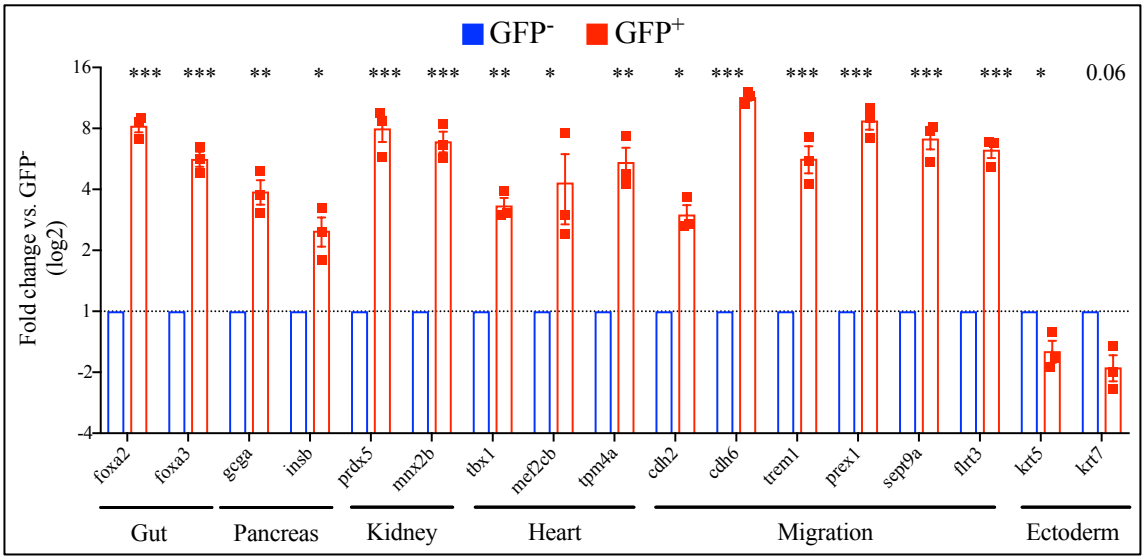


Figure 5.53 Validation of *sox17:GFP* RNA-seq results using RT-qPCR. Relative fold changes expressed as $\log_2$ of each gene in GFP^+ cells normalised to the expression in GFP^- cells (n=3). Student t-test *p*-values from RT-qPCR analysis ranged from 0.05 to $1.02E^{-05}$.

RT-qPCR certified the enrichment in GFP^+ cells for gut markers (*foxa2*, *foxa3*), pancreatic markers (*gcga*, *insb*), kidney markers (*prdx5* and *mnx2b*), heart formation (*tbx1*, *mef2cb*, *tpm4a*) and cadherin and GTPase dependant proteins (*cdh2*, *cdh6*, *trem1*, *prex1, sept9a* and *flrt3*). In addition, GFP^- cells were enriched for ectodermal markers *krt5* and *krt7*. The trend of differential expression of these genes was consistent with the RNA-seq data I generated and supported by the information in the literature.

As a side note, in Chapter 4, I argued the supremacy of TaqMan over SYBR chemistry in detecting gene expression in sorted cells via RT-qPCR. The results reported here were achieved with SYBR. It should be noted the validated genes were enriched in GFP^+ cells. Both

TaqMan and SYBR were adequate in detecting endodermal genes in the GFP^+ population, however SYBR was less accurate in revealing the enrichment of mesodermal/ectodermal genes in the GFP^- population. Here, I used *krt5* and *krt7* to show their selective enrichment in the GFP^- population; the differential expression of these genes was barely statistically significant ($p = 0.047$ and 0.06, respectively).

5.7.5 Transcriptome of non leaky and leaky embryos

The development and progression of endoderm formation is a complicated process where multiple factor coregulate fate outcome. I identified multiple genes that were differentially expressed during gastrulation when comparing leaky to non leaky embryos (*sox17*, *sox32*, *mixl1*, *myf5)* from the RT-qPCR results described in the Chapter 4. I therefore hypothesized that many other genes could be potentially affected in leaky embryos. However, identifying these genes by conventional methods such as serial analysis of gene expression would have been time intensive and not systematic. Transcriptomics have provided promise for massive gene transcript analysis therefore, as the next step towards obtaining an overview of the changes in leaky embryos, I prepared a single replicate library from sorted cells of leaky embryos (both GFP^- and GFP^+) and proceeded to compare gene expression levels to the baseline gene expression in non leaky embryos. I then assessed the overall gene expression levels comparing GFP populations of sorted cells (GFP^+ versus GFP^- cells) in leaky embryos and from the comparison of the transcriptomic profile, no difference was detectable in these populations. In addition, the transcriptomic profile of leaky GFP^+ cells resembled the transcriptomic signature of both leaky GFP^- and non leaky GFP^- (Figure 5.54). The PCA clearly separates GFP^- (red circles, left side) from GFP^+ (red triangles, right side) in non leaky embryos, whereas both leaky GFP^- (green circle) and leaky GFP^+ (green triangle) cluster together on the left side, closely positioned to non leaky GFP^- (red circles).

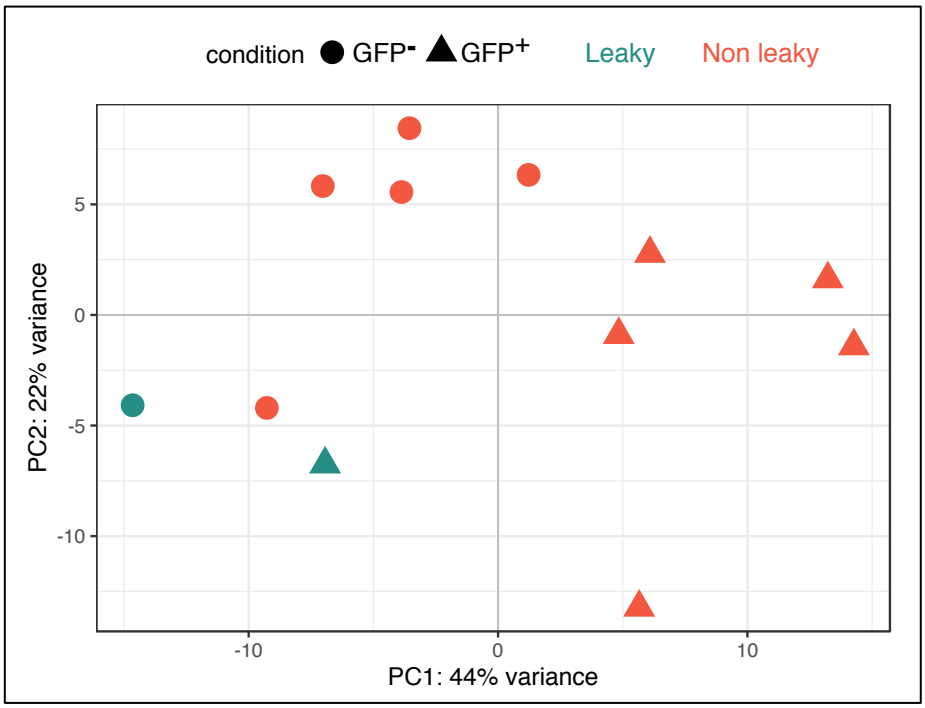


Figure 5.54 PCA plot for non leaky and leaky embryo. 10 libraries (5 GFP$^-$ and 5 GFP$^+$ cells) belonged to non leaky samples whereas 2 libraries (1 GFP$^-$ and 1 GFP$^+$) were collected from leaky embryos. In the latter, no separation between GFP$^-$ and GFP$^+$ libraries was visible (green circle and green triangle). The non leaky GFP$^+$ libraries clustered together with the GFP$^-$ libraries indifferently from leaky and non leaky embryos. The colours indicate the type of embryos: non leaky – green; leaky - red. The shapes indicate the type of population: GFP$^-$ – circle; GFP$^+$ – triangle.

The clustering analysis results were similar to those observed using RT-qPCR, where only the top GFP$^+$ cells showed endodermal marker enrichment, and cells with intermediate levels of GFP expression contained a mix of multiple germ layer cells. Consequently, the transcriptomic profiles generated by the RNA-seq analysis support the conclusions obtained by RT-qPCR and shown in Figure 5.55, whereby the leaky embryo transcriptomic signatures for both GFP$^-$ and GFP$^+$ (green tab, left side) closely match the transcriptomic signature of non leaky GFP$^-$ (orange tab, left side).

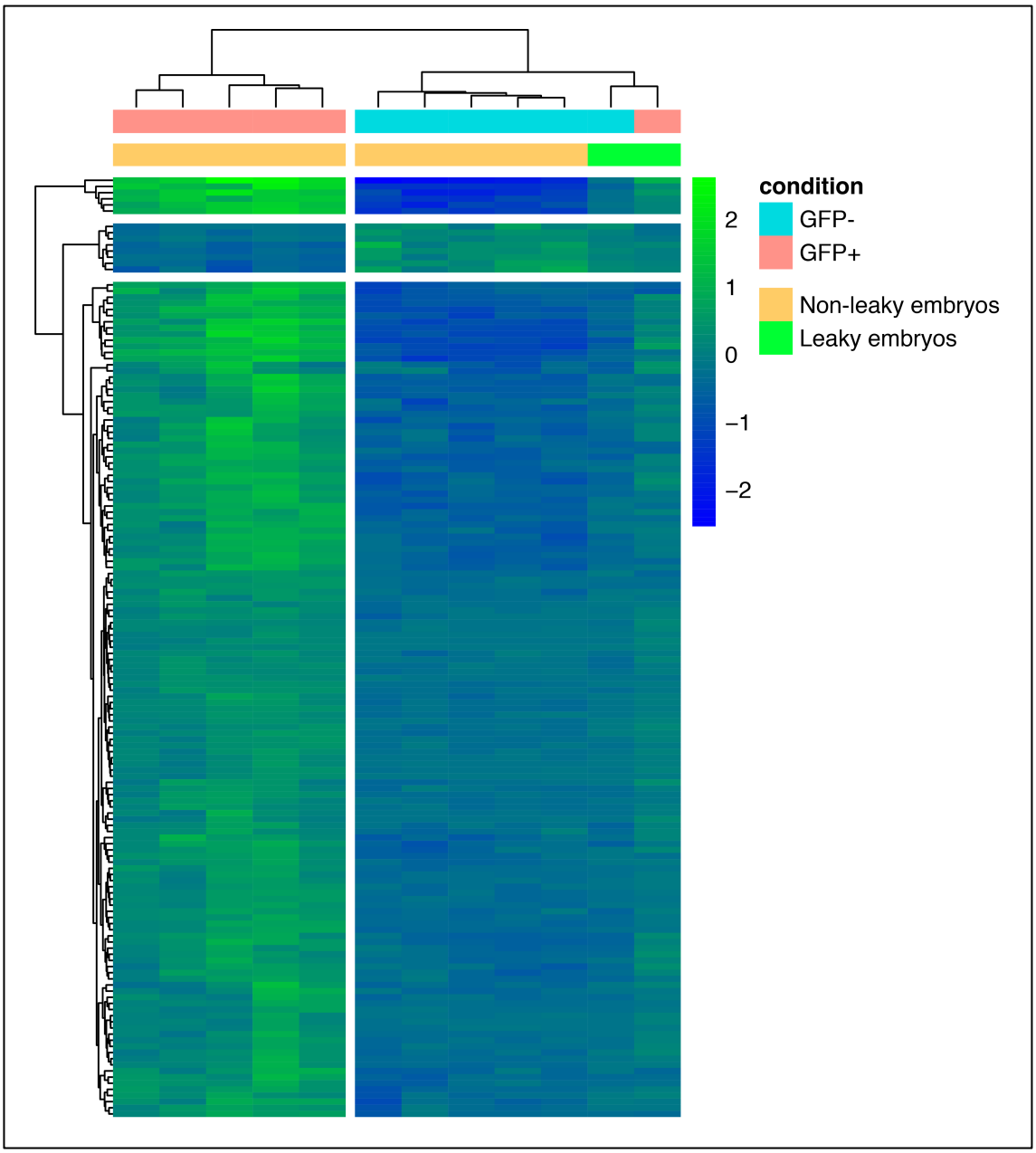


Figure 5.55 Gene clustering analysis of leaky and non leaky embryos. Clustering the transcriptomic data of GFP^{-} and GFP^{+} cells showed that leaky GFP^{+} had a similar gene expression levels to non leaky and leaky GFP^{-}. The leaky samples (bright green) were the sister cluster of non leaky GFP^{-} samples (yellow) on the right side of the plot.

5.8 Validation of endoderm specific genes and single cells RNA-seq

I then asked whether information on the spatial domains of genes in my list of DEGs was available from previous studies as well as whether information about their expression had been compiled on ZFIN.

For example, I was able to find the expression of *hkdc1* (a shared gene of "*mixl1*$^{-/-}$ 5.25 hpf", "*sox32*$^{-/-}$ 5.25 hpf" and "*sox32*$^{-/-}$ *9*.00 hpf"), *pck2* ("*sox32*$^{-/-}$ 5.25 hpf" and "*sox32*$^{-/-}$ 9.00 hpf"). Amongst the 6 common elements of "*sox32*$^{-/-}$ 5.25 hpf", "*sox32*$^{-/-}$ 9.00 hpf" and "*sox17:GFP* 9.00 hpf" datasets, information on the genes *tfa*, *slc43a2b* and *phospho1* was also available. No gene expression pattern data were available for *cdh6*. All these genes were found to be expressed in the YSL and their spatial domains can be compared to that of *sox32* in Figure 5.56.

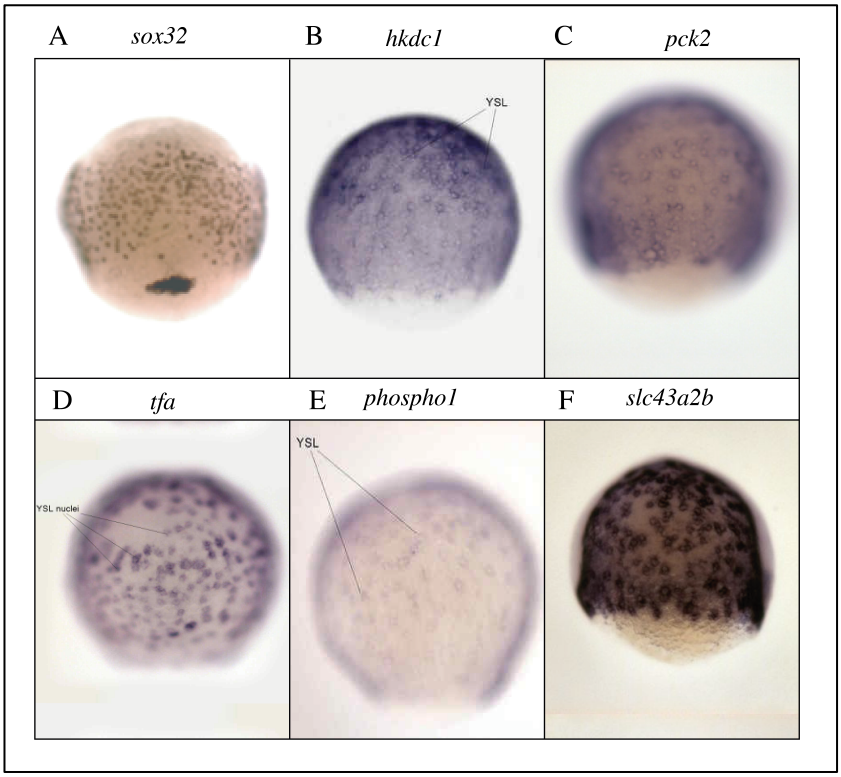


Figure 5.56 Spatial expression domain data from ZFIN. Expression of the indicated genes in WT embryos was visualised by downloading *in situ* hybridization information from ZFIN and comparing them to the *sox32* spatial expression domain.

For other genes of interest, once the genes were validated by RT-qPCR, I proceeded to design new probes and test their expression in WT embryos by *in situ* hybridization. Figure 5.57, shows the spatial expression domains of *cdh6 (*a marker of pancreatic and interrenal primordium), *prdx5* (expressed in cells in the pronephric ducts and gut) and *txn* (a marker of pharyngeal arches and gut) supporting the DEG analysis.

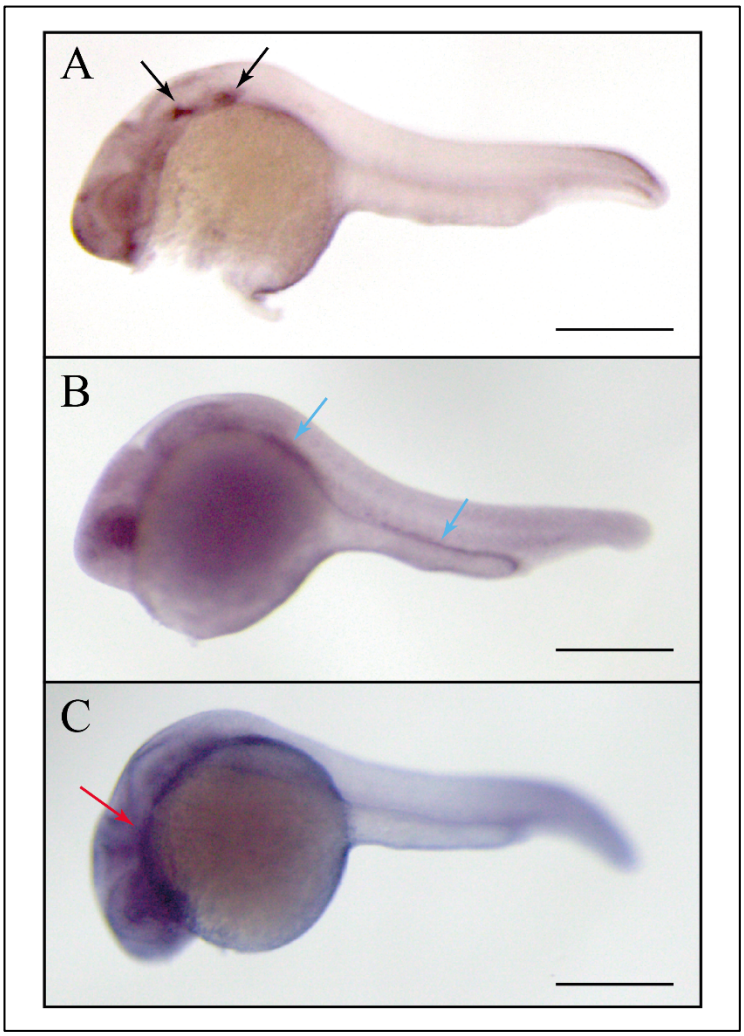


Figure 5.57 WISH was used to validate differentially expressed genes identified from RNA sequencing experiments. Expression of *cdh6* (**A**), *prdx5* (**B**) and *txn* (**C**) in 24 hpf embryos. Black arrows indicate liver and pancreas primordial respectively; blue arrows indicate gut and pronephric duct respectively and red arrow indicate the pharyngeal arches. Lateral views, anterior to the left. Scale bar: 300 µm.

I then compared the results of my RNA-seq analysis with published data, in particular to the recent single cell RNA-seq datasets (Farrell et al., 2018; Wagner et al., 2018). The information in these datasets matched the information discovered in my analysis in the *sox32* and *mixl1* mutants, for example, *txn* and *met* were clustered in the pharyngeal endodermal trajectory whilst *cdh6*, *prdx5* and *flr3* were grouped both in the pharyngeal and the hepatopancreatic trajectory (Figure 5.58).

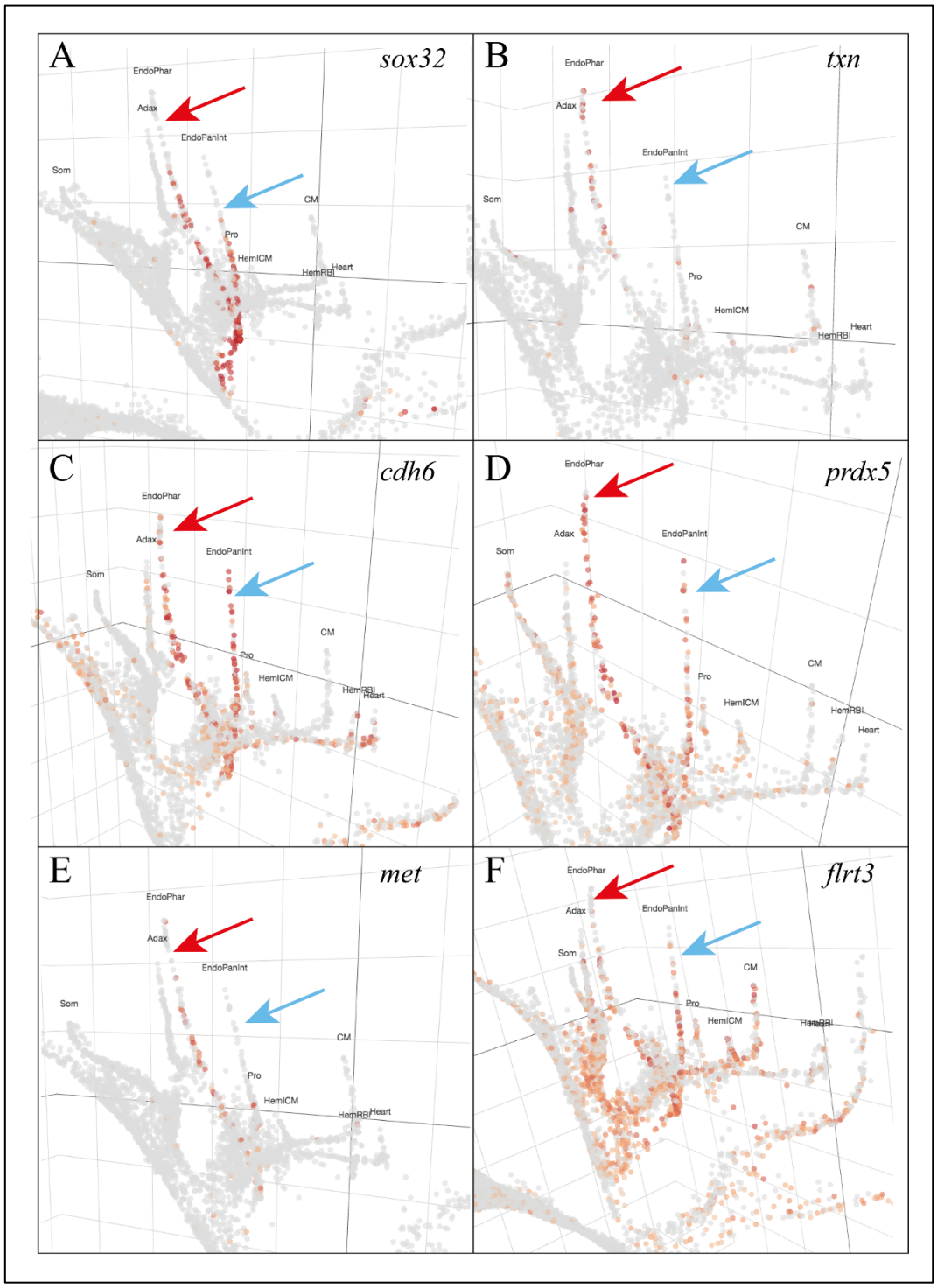


Figure 5.58 Single cell pseudotime trajectory trees reveal the developmental trajectories for endodermal genes. The transcriptional trajectories for *sox32* (**A**), *txn* (**B**), *cdh6* (**C**), *prdx5* (**D**), *met* (**E**) and *flr3* (**F**) are reported. The developmental trees describe the specification fate of cells where the highlighted genes are expressed. Expression of *sox32*, the master regulator of endoderm formation is observable in both the pharyngeal (red arrow) and the hepatopancreatic trajectory (blue arrow). Similar expression patterns were observable for *cdh6*, *prdx5* and *flr3*. *txn* and *met* were detected only the pharyngeal endodermal trajectory. Data and trajectory trees taken from (Farrell et al., 2018).

Taken together, the data showed that the differential expression patterns shown in my dataset were highly consistent between RT-qPCR, WISH and with the patterns demonstrated by these single omics techniques.

5.9 Technical validation - genotyping for RNA-seq

Global gene expression comparative analyses are a relevant tool to detect new genes underlying an observed phenotype; phenotypic variances are often not explained by a single gene but by the combination of expression of genes in a cohort. Such genome wide analyses can typically monitor changes in transcript abundance between experimental and control samples. In order to collect my experimental samples for RNA-seq, I needed to genotype embryos from both mutant zebrafish lines, *sox32*$^{-/-}$ and *mixl1*$^{-/-}$. Genotyping of embryos from both lines proved to be challenging.

The *sox32* mutant line was purchased from EZRC which shipped 30 or more embryos following *in vitro* fertilization of WT eggs. Once the adult fish were old enough, I fin clipped them to identify which fish carried the desired trait. Following guidelines, genotyping of the mutant was based on the Restriction Fragment Length Polymorphism method (RFLP) (Botstein et al., 1980). The allele contains a single T>G point mutation that introduces a premature stop codon at residue 170 of the Sox32 protein leading to a truncation of the protein shortly after the HMG domain (Dickmeis et al., 2001). In addition, this single point mutation also creates a site recognized by the BfaI restriction enzyme.

In the RFLP assay, the *sox32* sequence was first PCR-amplified and then the resultant PCR product was digested by the BfaI restriction enzyme; the presence of the mutation was determined by resolving the fragments on a 2% agarose gel and presence of the mutation determined by observing the resulting restriction pattern: WT only one band, mutant 2 bands. The use of this technique, although functional, was impractical for determining the genotypes of large numbers of fish, therefore to facilitate the detection and rapid genotypic analysis of *sox32*$^{-/-}$ carriers, I exploited high resolution melting (HRM) curve analysis (Parant et al., 2009). HRM methodology is powerful, rapid, high-throughput and specific for genotyping single nucleotide polymorphisms (SNPs) in a large number of samples and can be used as an alternative approach to direct DNA sequencing for the detection of SNPs. It is based on the generation of different melting curve profiles due to the presence of sequence variation in the

double-stranded DNA, as a single nucleotide change causes a shift in melting temperature. HRM analysis is being increasingly used for gene scanning because it is simple, cost effective, sensitive and relatively fast (Xing et al., 2014).

As shown in Figure 5.59, WT embryos showed only one peak in the melting curve whereas heterozygous carries showed a double 'bump'. Fish that were genotyped using RFLP matched the HRM methods, confirming the validity of this approach; therefore the heterozygous mutant sequence TAT/TAG and WT TAT/TAT can be distinguished by HRM analysis.

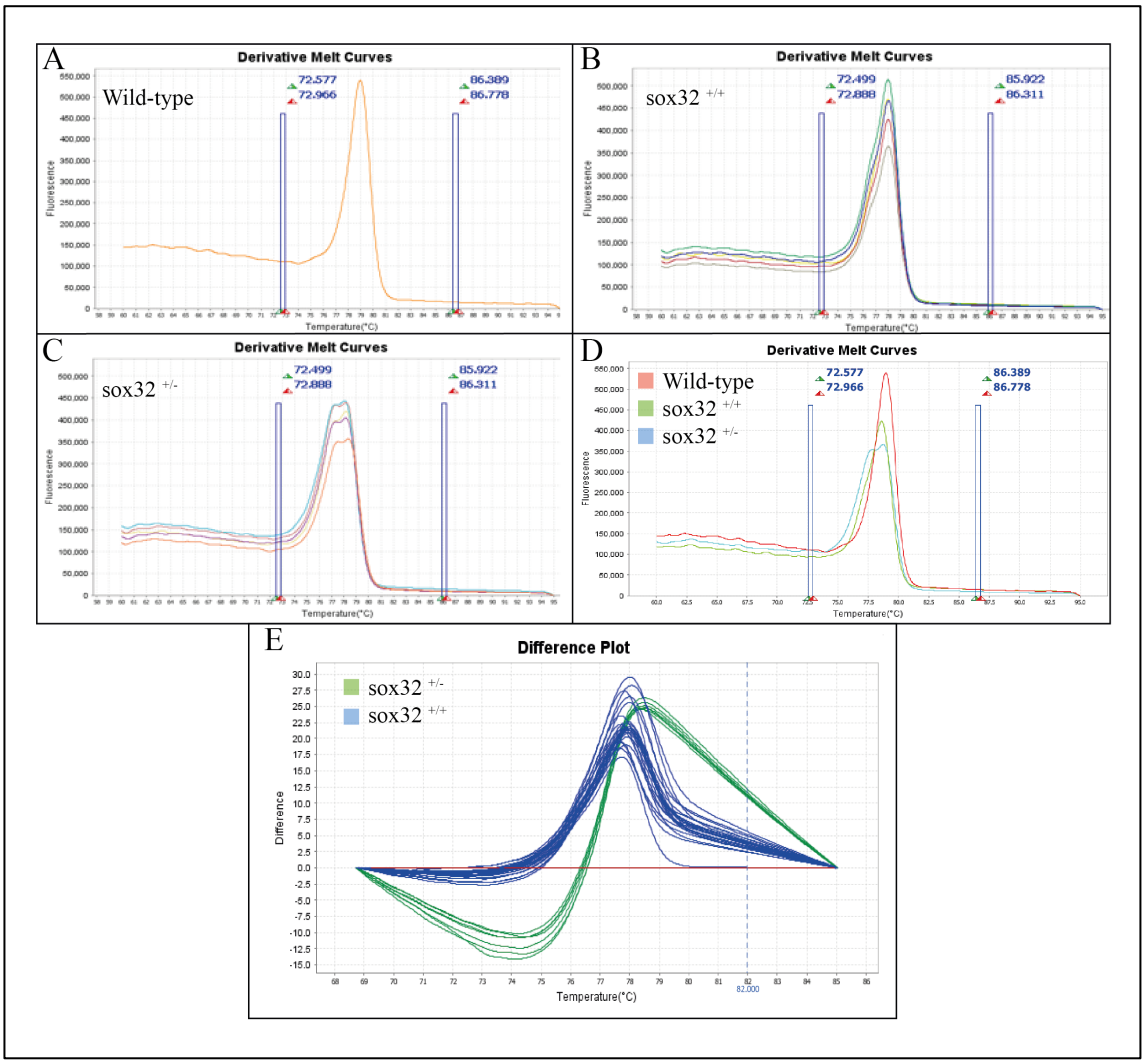


Figure 5.59 High resolution melting curve analysis can efficiently detect the *sox32* mutation in unidentified fish. HRM analysis of gDNA from fin clipped adult zebrafish shipped from ZIRC. HRM curves for a wildtype fish(**A**), for a *sox32* WT sibling (**B**) and a *sox32* heterozygote sibling (**C**). A total of 5 curves are presented to demonstrate the reproducibility of the assay and its ability to discriminate multiple curves (replicate curves are coloured). Note that curves in (C) have a double 'bump' at 75°C and 77°C. (**D**) Overlapping of (A), (B) and (C). (**D**) The HRM curves were compared with the *sox32* WT sibling curves as

the baseline. The 2 melting domains are evident; the first between 69 and 75°C and the second between 75 and 81°C.

RNA-seq analysis is typically based upon quantitative assessment of transcript abundance which is then compared between a WT and a mutant sample. The use of heterozygotes mutants for *sox32* was impractical for my RNA-seq experiments in which I wished to perform comparative transcriptome analysis, because the *sox32* heterozygote displays evidence of gene dosage compensation, as shown by RT-qPCR data of the downstream genes *sox17* and *foxa2*. RT-qPCR of these genes showed no difference in transcript abundance between WT and heterozygote, yet these genes are severely downregulated in the homozygotes (Figure 5.60).

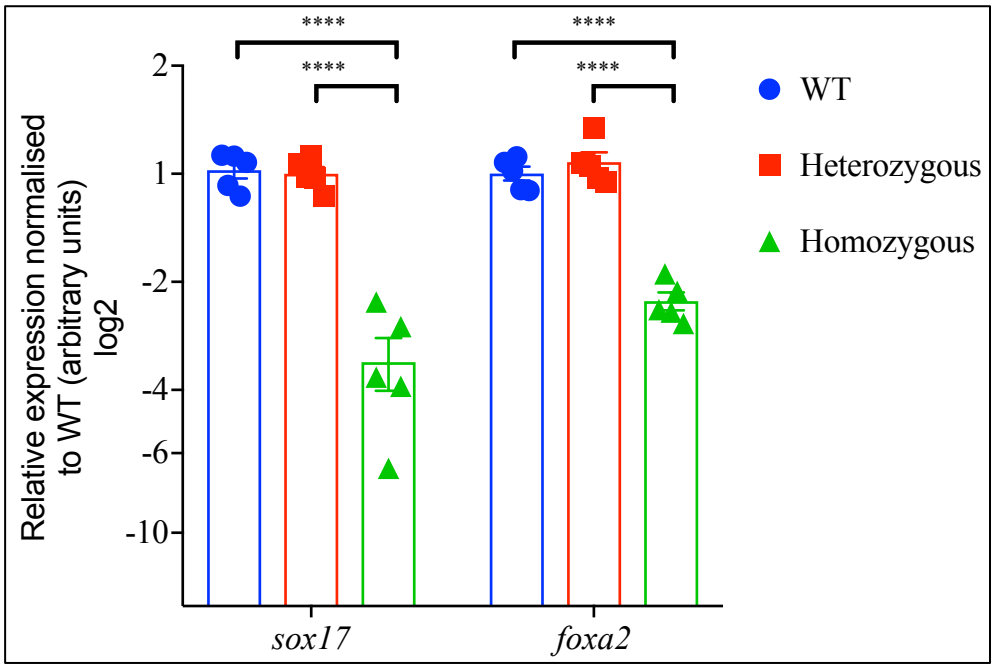


Figure 5.60 Relative expression of genes downstream of *sox32* in *sox32* mutants at 9.00 hpf. Downregulation of both *sox17* and *foxa2* is only observed in homozygous fish. Each point is an embryo. Data are represented as mean ± SEM (n = 5) and fold change is displayed relative to WT expression. One-way Anova (with Tukey post-hoc test) was used to assess statistical differences. ****$p \leq 0.001$.

I therefore started to use incrosses from *sox32* heterozygotes, which yield 1/4 *sox32* homozygous offspring. Single embryo genotyping was then selected to obtain only homozygote embryos, in order to fully understand how the Sox32 non functional protein was impairing the development and formation of endoderm in zebrafish. No phenotypical defects are observable in these mutants during gastrulation and I was therefore unable to determine the genotype of live embryos. As homozygous, heterozygous and WT siblings are indistinguishable during gastrulation, I first sorted the embryos at 24 hpf, using a previously

identified marker (*myl7*) then tested whether HRM could be applied to identify the genotype of individual Sox32 embryos resulting from heterozygous incrosses.

Observation of WISH using a probe for *myosin light chain 7 (myl7*) in 24 hpf embryos revealed that 73% of embryos from a single heterozygote incross exhibited no phenotypic heart defect and 27% of embryos presented cardia bifida (Figure 5.61). Among the 73%, 20% were WT and 53% were heterozygous for the *sox32* mutation. These numbers were therefore close to the expected Mendelian ratio. These embryos were then subject to HRM analysis.

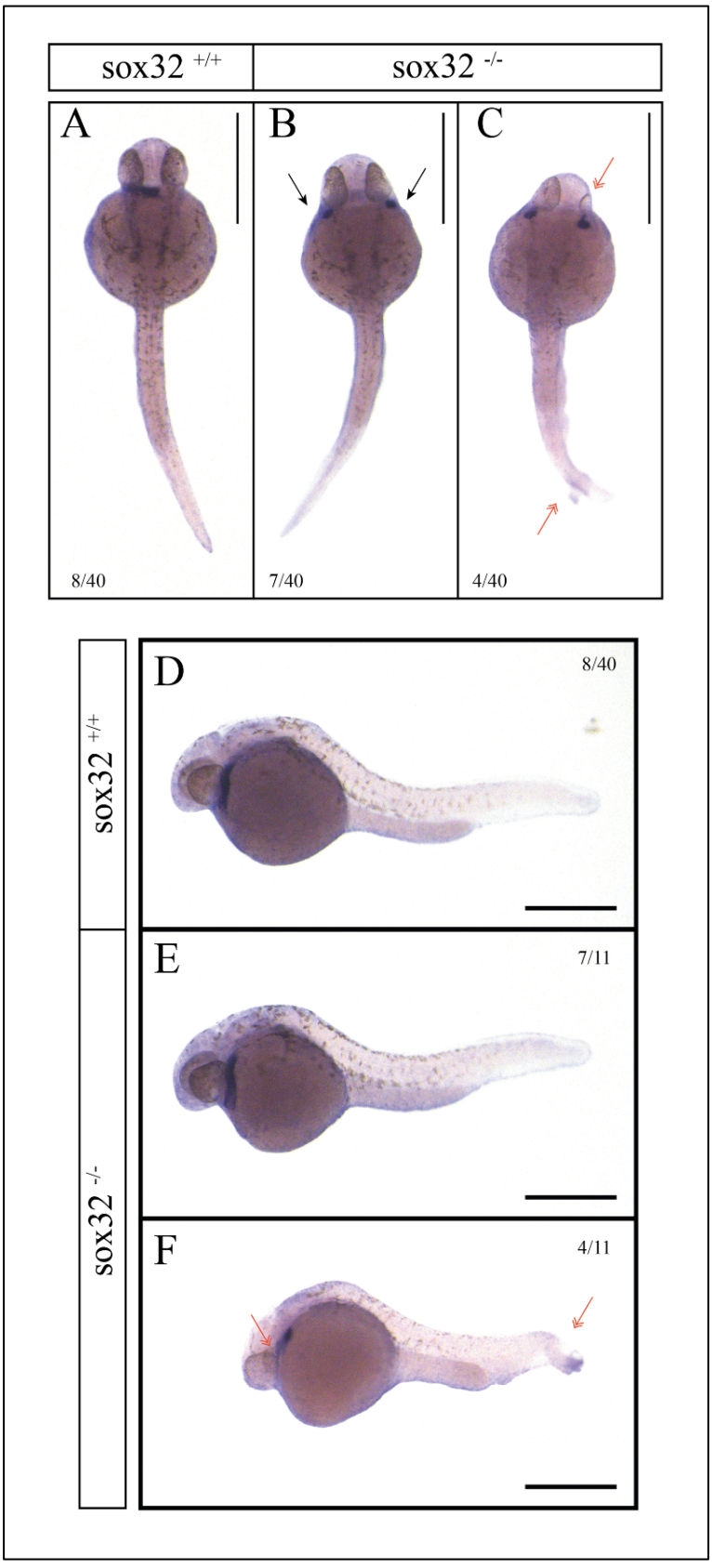


Figure 5.61 Heart defects visualised via WISH for *myl7* at 24 hpf. (**A** and **D**) WT sibling embryo with normal heart looping. (**B**, **C**, **E** and **F**) Homozygous mutant sibling embryos with cardia bifida. Black arrows depict bilateral hearts, red arrows depict eye malformations and short tails (split or branched). (A-C) are dorsal

views; (D-F) are lateral views, anterior to the left. The number of embryos with the indicated expression pattern among the total examined in 2 biological replicates is shown. Scale bar represents 100 µm.

I proceeded to try HRM to genotype the progeny derived from the heterozygous parental cross. I tested 3 different sets of primers in the search for optimal parameters for reliable detection of the homozygous allele. However, I was unable to unequivocally identify homozygotes from WT (Figure 5.62). HRM failed to recognise the T to G transversion (TAT → TA*G) of the *sox32* allele from the WT allele.

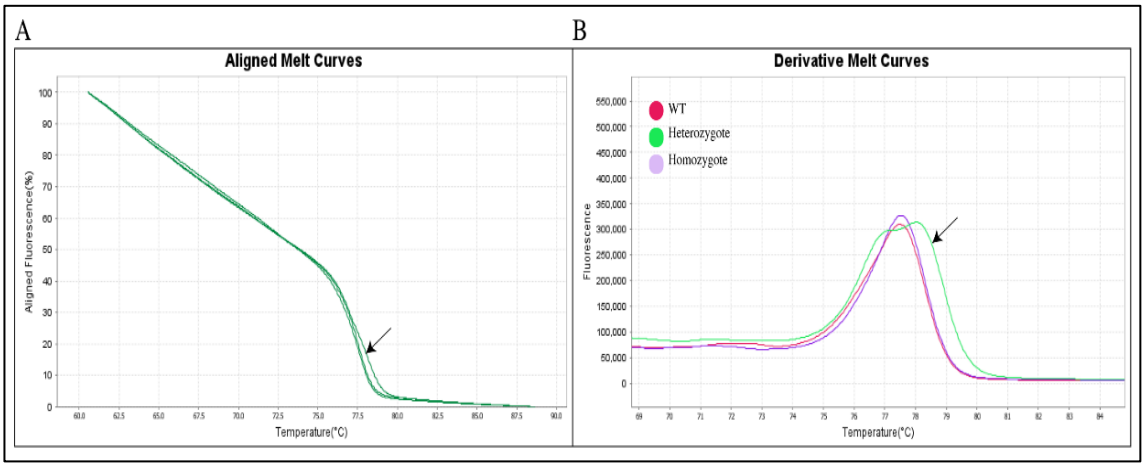


Figure 5.62 High resolution melting curve analysis cannot detect homozygotes in *sox32*$^{-/-}$ fish. (**A**) Aligned HRM curves for a *sox32* WT sibling, *sox32* hom sibling (**B**) and a *sox32* heterozygote sibling (black arrow). (**B**) Derived melting curve for the 3 samples. Note that heterozygoteus siblings have a double 'bump' at 76 °C and 79 °C. No difference in WT and homozygotes melting domains are evident.

In conclusion, to collect the embryos needed for the RNA-seq experiment, I needed to individually sequence gDNA extracted from single embryos (Figure 5.63), saving the total RNA collected from the embryos for RNA-seq library preparation. Once I had identified 9 x WT and 9 x homozygotes siblings, 3 were pooled in triplicate to generate 3 biological replicates for each condition. This process was repeated for both 2 time points; 5.25 and 9.00 hpf.

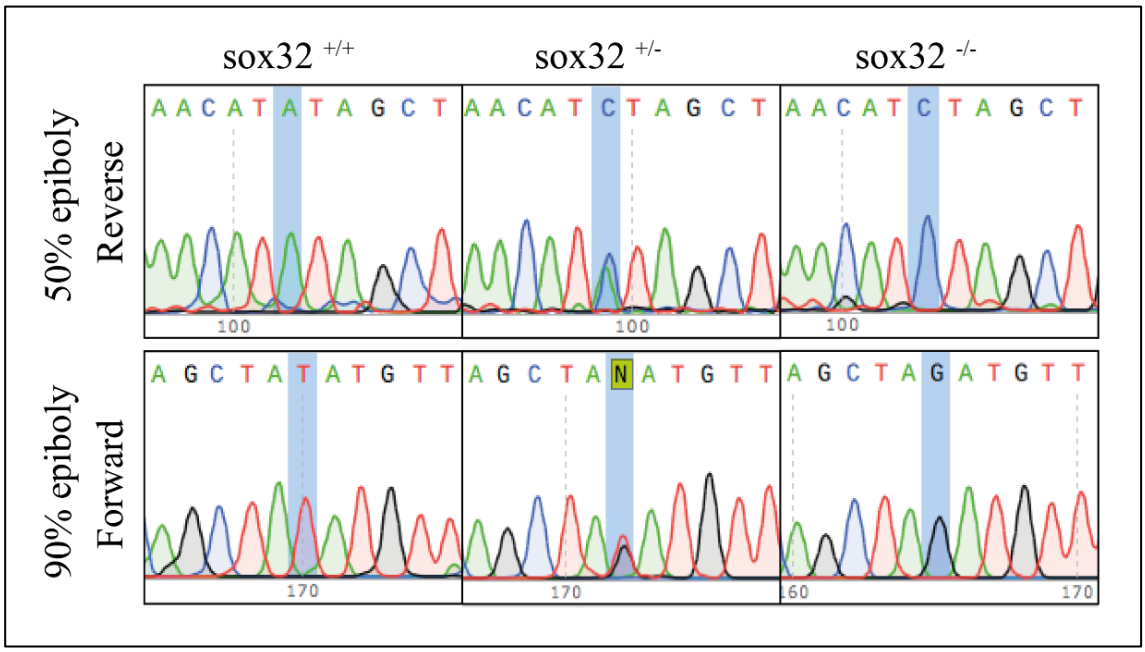


Figure 5.63 Electropherogram for DNA sequence analysis of *sox32* mutant. The mutant allele presents a T>G point mutation (blue shading) that introduces a premature stop codon in the protein. Examples of representative electropherogram traces from gDNA extracted from embryos at 5.25 hpf (50% epiboly) and 9.00 hpf (90% epiboly). To be sure not to mix the stages, samples from the 2 time points have been both amplified and sequenced with different primers. All samples extracted from embryos at 5.25 hpf (top panel) have been sequenced with a primer on the reverse strand, with the WT sibling have an A and the mutant a C. All the samples at 9.00 hpf (90% epiboly), (bottom lane) were sequenced using a forward primer and have a T in the WT and a G in the mutant. Note that heterozygotes embryos (middle panels) show double peaks.

Although rapidly identifying a sufficient number of *sox32* homozygous mutant embryos was a challenge, as they are morphologically indistinguishable from their siblings during the gastrulation process, I am convinced that once the challenge of genotyping was optimised, the mutant approach gave better results than possibly using a Sox32 knocked down with morpholino, as no stress response protein or innate immune response genes were detected in the differential gene expression. RNA-seq with morphants is a viable option in absence of mutant line but requires titration and optimization of concentration of morpholino to not only phenocopy the mutant (*sox32*) defects but also obtain robust results. In *Xenopus* both longer incubation time with lower temperatures and optimization of morpholino dosage alleviate, but not eliminate, these side effects in the morphants (Gentsch et al., 2018; Stainier et al., 2017).

5.10 Technical validation – validating *mixl1* RNA-seq data with RT-qPCR

In order to validate the results of the RNA-seq, I collected new biological replicates for both Mixl1 and Sox32 mutants (and WT siblings) and genotyped the embryos as described

above. In the case of the *sox17:GFP* line, this involved 3 additional rounds of FAC sorting to obtain 3 biological replicates. Total RNA from the samples was then converted to cDNA and subject to RT-qPCR analysis to test whether the differential gene expression levels identified in the RNA-seq data were accurate.

At this point, genotyping the *sox32* mutants was relatively straightforward as I had already optimised the protocol in order to prepare the RNA-seq libraries. Between the time of sequencing the Mixl1 libraries and completing the bioinformatics analysis, all the females from the original 10 *mixl1* homozygous fish that were shipped from the D.M. lab in Austria had died. A new generation was derived via artificial insemination of WT fish; homozygous males were sacrificed, gonads harvested and sperm collected. Once the fish grew, fin clipping and genotyping of heterozygotes carrier was done. I then obtained embryos resulting from the incross of heterozygous carriers and I tried to identify homozygous siblings using Sanger sequencing.

The Mixl1 mutation introduces a T to A transversion (TAT → TAA) in the coding sequence of the gene which introduces a premature stop codon at amino acid residue and a truncation in the homeodomain, the site of the protein that bind to the regulatory regions of Mixl1 target genes, thus Mixl1 can no longer bind to its targets. This point mutation also creates a new MseI restriction site that is not present in the WT.

In the original paper, (Kikuchi et al., 2000) used the PCR-RFLP protocol on embryos at 30 hpf to score the genotype and match it to the observable phenotypes. However, amplified PCR fragments using the same conditions and primers as described in the paper did not produce reliable results. I used enzymes from both Roche and NEB but MseI did not efficiently work in the PCR buffer displaying a cutting activity of less than 25%, resulting in a lot of false positives. Addition of more enzyme units and/or addition of the appropriate enzymatic buffer to the PCR mix did not help to improve the ability of the enzyme to cleave the PCR product. However, I found that the same assay worked when the PCR product was first purified by either PCR column kit or by ethanol precipitation and then incubated for 12-16 hrs with MseI.

As described earlier, HRM analysis has rapidly become an important gene scanning technique as it allows genotyping without the need for costly enzymatic digestion. However, this methodology has its limitations, as shown by the failure to detect the single nucleotide

substitution in the *sox32* mutant (data not shown). Similarly, the *mixl1* homozygous sequence was indistinguishable from the WT sequence by HRM analysis (data not shown).

To identify the *mixl1* mutation, I decided to amplify the genomic locus and sequence it, as HRM genotyping was not able to distinguish between the 3 genotypes (WT, heterozygotes and homozygotes). Although the RFLP assay with PCR purification worked effectively, it proved to be more time consuming and expensive than directly sequencing the samples, as described previously for the *sox32* mutation.

5.11 Discussion

In this chapter I applied RNA-seq technology to compare WT and mutant embryos for 2 genes, *sox32* and *mixl1*. Lack of activity of these proteins leads to well characterised and defined endodermal defects, and my aim was to contribute to the understanding of the underlying transcriptional network regulated by Sox32 and Mixl1 respectively. I additionally combined RNA-seq and FACS to isolate a distinct endodermal GFP^+ cell population from a heterogeneous sample - dissociated *sox17:GFP* embryos - where GFP^+ cells represented endodermal cells as GFP expression is under the control of the *sox17* promoter.

Many of the techniques I described have been broadly applied to other tissues or model systems in literature, but I applied them for the first time in zebrafish embryos to better understand the gene regulatory network governing endoderm development. Over the last few years, RNA-seq has become a powerful tool to profile gene expression of any given system, being more specific, with a larger dynamic range and higher sensitivity, and requiring less input material than the previously used microarray technology. Additionally, knowledge of the transcriptome is not a requirement for RNA-seq and it is therefore applicable for the study of novel transcripts and isoform splicing. In essence, this technique has superseded and replaced microarray technology.

Transcriptomic studies using both microarray and RNA-seq have been positively incorporated in zebrafish research to explore and detect new genes related to biological function and cell types. The start of the ‘omics’ era has contributed significantly to the study of GRNs. An increasing catalogue of sequenced genomes and the availability of simpler and cheaper protocols to do genome-wide techniques has promoted an explosion of exploratory studies that have not only amplified the documentation of unknown developmental players but

have also confirmed previously known TFs (Lowe et al., 2017; Rafiq et al., 2014; Simões-Costa and Bronner, 2015; Simoes-Costa et al., 2014; Williams et al., 2018). Despite all the advantages that RNA-seq offers, it still has some technical difficulties to consider and bias to recognise and mitigate. I have highlighted the importance of RNA quality, as input with degraded RNA results in less usable libraries, exemplified by the *mixl1* mutant libraries from Austria. I demonstrated the importance of minimizing the PCR amplification bias, specifically that starting with higher amounts of RNA input was favourable, and that a higher number of PCR cycles correlated generally with a higher number of duplicated reads, which should be avoided. Despite these limitations and technical problems, I prepared libraries and performed RNA-seq for *sox32* and *mixl1* mutants and compared their respective transcriptomes to those of WT embryos. I also prepared libraries and performed RNA-seq on cells sorted from the *sox17:GFP* line, enabling me to compare the endodermal transcriptome with that of ectodermal/mesodermal cells.

In order to evaluate differential gene expression in my experimental conditions, I adapted the main frameworks and methods from the ENCODE and DANIO-CODE bioinformatics pipeline and characterised the global scenario of the regulatory relationships between Sox32 and Mixl1 regulators and their targets. Taking into account only genes whose expression levels differed from the control by at least FC > 1 and FDR < 0.01/0.05, I was able to identify 2075 downregulated genes and 1996 upregulated genes in the *mixl1* mutant at 5.25 hpf, 252 genes downregulated and 192 upregulated in the *sox32* mutant at 5.25 hpf and 189 downregulated and 108 upregulated genes at 9.00 hpf. In addition, the analysis of sorted *sox17:GFP* cells revealed 349 genes enriched in endodermal cells of which a significant proportion were novel.

I first identified several genes that were differentially and significantly regulated in each experiment and as a next step, I organised the data to see which genes were overlapping in all conditions. I asked whether a pattern emerged and I then focused on verifying genes whose change in expression level was shared among all the conditions, using both RT-qPCR and *in situ* hybridization. The choice of focusing on genes that all data sets had in common was made in order to find new interactive nodes to add to the GRN (see Chapter 6) and not only to characterise the mutant signature *per se*. The rationale and hypothesis underpinning my decisions were that a global regulatory network was embedded within regulators (Sox32/Mixl1) and targets with high interacting affinities, which could be learned from

transcriptomic data, and that the details of individual regulatory relationships could be validated by further experiments.

Notably, analysing both datasets at 5.25 hpf allowed me to identify new promising genes that could explain the different regulatory circuits downstream of *mixl1* and *sox32* governing fate decisions in the mesendodermal cell population; there were both common and distinct mechanisms underlying the action of Mixl1 and Sox32 in each case. At 5.25 hpf my results for both TFs yielded genes that were expressed primarily in the YSL and margin and genes belonging to the Nodal signalling pathway. 2 particularly interesting genes observed to be upregulated in the *sox32*$^{-/-}$ mutant were *nanog* and *mxtx2* which are directly controlled by maternal factors such as Eomes (Bruce et al., 2005; Du et al., 2012; Xu et al., 2014). These genes are linked intrinsically to establish autoregulation loops of Nodal ligands in the YSL and in the margin from the midblastula stage**.** Nodals in return start a cascade of signalling that ultimately leads to the foundation of endodermal and mesodermal territories as confirmed by studies in several other mutant lines (*ndr1*, *ndr2*, *gfd3*, *lft1/2, acvr1ba*) all of which show different degrees of severity of endodermal and mesodermal defects (Bisgrove et al., 2017; Chen and Schier, 2002; David and Rosa, 2001; Dougan et al., 2003; Feldman et al., 2000; Montague and Schier, 2017; Peyrieras et al., 1998; Rogers et al., 2017; Schier et al., 1997). A recent study (Veil et al., 2018) also showed how maternal deficient embryos of *nanog* and *mxtx2* fail to survive after gastrulation and have delayed and lower expression levels of early endoderm specifying genes *sox32* and *mixl1* compared to WT embryos. *gata5* was also downregulated by lack of *nanog* and *mxtx2* activity. These results can be expanded upon with my study of *sox32*$^{-/-}$ at 5.25 hpf, where, in the absence of Sox32 functional protein, *nanog* and *mxtx2* expression was upregulated. It is therefore possible that in the wildtype embryos, Sox32 once activated acts as a direct repressor of these early Nodal inducers.

This role of Sox32 can be explained as the mesendodermal circuit shutting down the previous module of Nodal regulation in the GRN (see Chapter 6). Shutting down and regulating the window of competency of Nodal signalling may be another role of Sox32 in the cells in the margin of the embryo. To support this speculation, other genes that regulate the Nodal domain were affected in *sox32*$^{-/-}$ and *mixl1*$^{-/-}$ mutants; in particular, *dusp4/6*. These genes have been associated with endoderm formation in zebrafish (Brown et al., 2008) and in particular Nodal induces short range *dusp4* expression within the first 2 cell tiers in the margin and simultaneously induces long range Fgf signalling via p-Erk which inhibits endoderm

specification and commits the more distant cells from the margin to mesodermal fate (van Boxtel et al., 2018). Thus, *dusp4* attenuates p-Erk levels (and hence Fgf signalling) close to the Nodal source and enhances the specification of endodermal progenitors. In addition, my data suggest that the expression of *dusp* genes is coordinated and positively reinforced by *mixl1* and *sox32.*

The activity of Sox32 and Mixl1 on multiple regulatory regions (some overlapping, some distinct) of genes such as *dusp4/6/27, lft2*, *nanog* and *mxtx2* elegantly integrates into the model of multiple feedback interactions and dynamic responses to internal and external signals and pathways, which ultimately provide cells with temporal and positional cues that direct their fate. Moreover, the analysis of the 5.25 hpf datasets showed how both TFs regulate common mesendoderm patterning genes (52 common genes). For example, *hkdc1* that is important in the insulin pathway (Yang et al., 2017), *notum1a* which blocks the Wnt/β-catenin signalling pathway (Flowers et al., 2012) and *pkd2* that plays a role in the propagation of Nodal signals and restricting left side specific expression of southpaw (*spaw*) (Schottenfeld et al., 2007). Furthermore, in *sox32*$^{-/-}$, RT-qPCR verified changes in gene expression for *dlc*, *tbxta*, *sp5l*, *her1* and *eve*. Interestingly, Nanog is necessary for correct spatial expression of the ventral specifying genes *bmp2b*, *vox* and *vent*, and the neural transcription factor *her3*. This is in line with the changes in expression observed in the *sox32*$^{-/-}$. Sox32 regulates *nanog* and disrupting this interaction could cascade downstream, for example by affecting *her1/3* gene expression. However, not all genes that varied in the *sox32*$^{-/-}$ embryos can be explained solely by the reduced expression of *nanog*. In particular, *tbxta* levels do not change in the MZ*nanog mutant (Veil et al., 2018).*

Although my data is generally in agreement with previously published studies, my results partially contradict reports earlier in the literature, where *sox32* has been described to activate endodermal and forerunner specific genes autonomously and to repress mesodermal specific genes (Aoki et al., 2002; Kikuchi et al., 2001). My results highlighted how Sox32 controls *tbxta* expression in the mutant, but according to the literature, if Sox32 was repressing *tbxta* its level should have been higher in the mutant and not the opposite. It is possible that i) other factors are regulating *tbxta* expression and higher expression levels in the WT compared to the mutant are not directly related to the role of Sox32 or ii) *sox32* and *tbxta* are involved in cross-regulatory interactions. In respect of the latter, I can speculate that it is the level/amount of transcripts in each cell that leads to either coexistence of the *sox32*/*tbxta* transcripts or

generation of mutually exclusive domains during gastrulation, but this relationship was not captured by whole embryo RNA-seq. scRNA-seq has started to elucidate the heterogeneity of cell transcriptomes during development, for example, Yuan et al. (2018) clustered cells' fates based on the presence of *gata5*$^{+}$, *sox17*$^{+}$ and *sox32*$^{+}$ for endoderm, and *gata5*$^{+}$, *sox17*$^{–}$ and *sox32*$^{–}$ for mesoderm. My whole embryo RNA-seq captured average changes in the transcriptome, therefore it is possible that with scRNA-seq a subpopulation of cells with different levels of *tbxta*/*sox32* in the margin could be distinguishable (*sox32*$^{+}$, *tbxta*$^{+}$, other TFs^{+} vs *sox32*$^{+}$, *tbxta*$^{-}$, other TFs^{+} vs *sox32*$^{-}$, *tbxta*$^{+}$, other TFs^{+}). More interestingly, my overall results showed how Sox32 target genes account for its role in both endoderm and mesoderm formation and these data thereby create an anchoring point to link the new information I am presenting to the *tbxta* GRN that was previously described (Morley et al., 2009).

Analysis of the datasets produced in this study has not only added new information regarding the function of both Mixl1 and Sox32 proteins but also supported and confirmed previous genetic knowledge. According to the literature, loss of *mixl1* results in decreased expression of *sox17*, *foxa2* and *foxa3* (Kikuchi et al., 2000), observations which were confirmed by my RNA-seq results. Additionally, loss of Sox32 has been reported to result in decreased *sox17* and *foxa2* expression (Alexander et al., 1999; Dickmeis et al., 2001; Kikuchi et al., 2001), which is again consistent with my RNA-seq results.

Interestingly, overlapping the 2 distinct transcriptomic signatures revealed a large number of genes whose modified expression levels are specific to either Mixl1 or Sox32. This is consistent with the hypothesis that different gene regulatory networks are hardwired in the genome and convey overlapping functions in regulating endoderm development. Sox32 occupies some endodermal *cis*-regulatory modules which, for example, regulate *sox17* and *foxa2* expression, while Mixl1 mediates regulatory element activity in some, but not all Sox32 *cis*-regulatory modules. However, at the molecular level, gene expression is affected differently in the 2 mutants, with Sox32 regulating more genes related to heart formation than Mixl1, which may explain why *sox32* mutants have a more severe endodermal and mesodermal defects.

The activities on the *cis*-regulatory modules, either as a concerted effort between both transcription factors, or specifically regulated by a single one of them, is what allows a specific developmental output; in other words, these genetic interactions impose a specific

developmental outcome on the cell. This aspect of endoderm development was apparent in the pharyngeal arch trajectory where I observed how Mixl1 exclusively controls genes such as *dlx3, fras1*, *sox9a*, and *prdm1a* while Sox32 exclusively controls *met, txn* and *irx7* (Talbot et al., 2010; Talbot et al., 2012; Yan et al., 2002). In addition, both TFs regulate the common pharyngeal genes *vwf*, *flrt3* and *ednraa*. *itga5*, a gene downstream of *prdm1a* was also found to be expressed in the GFP^+ enriched population from the *sox17:GFP* line (LaMonica et al., 2015), whereas genes *fgfrl1b*, *fgfr2*, *fgfr3* and *col9a2* were exclusive enriched for GFP^+ but not in the mutants (Hall et al., 2006). These results highlight the role that TF networks play in collaboratively regulating endoderm development and also that gene regulatory networks are divided into functional subcircuits.

Interestingly, because of the way the library was prepared, the *mixl1* dataset highlighted the largely underexplored role of non-coding RNAs in regulating gene expression, as most of the DEGs were comprised of non-coding RNA and other unannotated genes that have yet to be characterised. Some papers have started to elucidate the emerging roles of non-coding RNAs in zebrafish, for example looking at the roles of miR-430 during zygotic genome activation (Lee et al., 2013), in the evolution of the Nodal signalling domain at the margin at the midblastula stage (van Boxtel et al., 2015) and in regulation of endoderm formation and L-R asymmetry by miR-92 targeting *gata5* (Li et al., 2011). Tackling the non-coding genome has uncovered non-coding RNA molecules that form part of the genetic regulation underlying specific cellular functions and are important for the regulation of endodermal cell fate decisions in cell lines (Hinton et al., 2010; Ishikawa et al., 2017; Ma et al., 2016; Yang et al., 2014); further exploration and experiments to assess function of non-coding RNA interference in zebrafish endoderm development is required.

In order to understand better the regulatory function of Sox32 during gastrulation I compared the 5.25 hpf to the 9.00 hpf datasets, this data mining showed how Sox32 not only keeps regulating itself throughout gastrulation but simultaneously directs a network of TFs, signalling and differentiation genes that ultimately work on 3 different levels: specification of endodermal fate, specification of mesodermal fate and control of cell migration.

Noteworthy, the common 33 downregulated genes at the 2 time points were linked to both Smad and Sox motifs; which strongly supported i) the importance of Nodal signalling through Smads in collaboratively orchestrating mesoderm and endoderm GRN and a potential role of Sox32 in directly controlling players of the Nodal pathway. Operating on multiple levels of

the signalling cascade, the systems can be used to establish the stable expression of mesendoderm transcription factors and finely tune Nodal signal both in space and time ii) Sox32 autoregulates itself, upholding the lock-on system operating in mesendodermal specification introduced by Chang et al. 2009.

As previously noted, the comparison between the transcriptomes at 9.00 hpf and 5.25 hpf in the *sox32* mutants clearly depicted the changing role and plasticity of the Sox32 network during development. The DEGs at 5.25 hpf were strongly associated with the YSL domain and induction of Nodal signalling, whereas at 9.00 hpf the DEGs were more closely linked to morphological movement, heart and endodermal development. At the end of gastrulation, Sox32 activated a cascade of new genes associated with late stage endodermal and mesodermal derivatives. These downstream genes, exemplified by *foxa2*, *met, aldh1a2,* and *jag1a,* are important in pancreas and liver formation, with respective mutants and/or morphants exhibiting a complete or partial loss of liver and pancreas duct lineage (Alexa et al., 2009; Anderson et al., 2013; Gao et al., 2008; Latimer and Jessen, 2008; Zecchin et al., 2005; Zhang et al., 2017). The role of Sox32 in DFC formation and differentiation has been described previously, with *sox32* mutants having fewer DFC cells, a defective KV and exhibiting L-R asymmetry defects (Alexander et al., 1999; Essner et al., 2005). My dataset not only captured changes in known genes which function in the DFC and KV such as *sox17* and *chd* (Aamar and Dawid, 2010), it also uncovered genes such *dnah9*, *spag6* and *foxj1a,* previously not reported to be under the control of Sox32 and critical in the establishment of L-R asymmetry in zebrafish (Chocron et al., 2007; Hellman et al., 2010; Tian et al., 2009).

Among target genes whose expression was significantly downregulated in *sox32* mutant were some involved in pronephric kidney development such as *peroxiredoxin5* (*prdx5*). The Peroxiredoxin family of proteins have been reported to play an important role in pronephros development and reduction in the expression of key gut developmental genes *vegT*, *pax6* and *sox17* in *Xenopus* (Chae et al., 2017; Peng et al., 2004) and zebrafish *prdx1* was recently identified as novel regulator of pronephros development (Chae et al., 2017). Here, I discovered *prdx5* which, according to both RT-qPCR and WISH, is expressed in the developing kidneys during zebrafish embryogenesis. Prdx5, being an antioxidant enzyme like Prdx1, catalyses the reduction of H_2O_2 and reduces cellular levels of reactive oxygen species (ROS). ROS impair primary cilia formation (Ji et al., 2018), possibly by harming cilia motility proteins such as the previously defined Spga9 and Lad1. These gene were both downregulated in *sox32* mutant at

5.25 hpf, which present clear ciliary defects in the Kupffer's vesicle, a ciliated organ of asymmetry; hence reduction in genes associated with normal cilia development suggest that Sox32 regulate multiple factors and the overall misexpression of them led to the phenotypic defects observed in the mutant.

An additional gene associated with kidney organogenesis and downregulated in *sox32* mutants at 9.00 hpf was the homeobox transcription factor *mnx2b,* which is required for pronephric tubule morphogenesis and function (Ott et al., 2016). 2 direct targets of the Mnx transcription factor, *irx1a* and *irx7,* both previously linked to kidney defects (Ott et al., 2016) were also detected in my analysis. These analyses reveal a novel interaction of Sox32 with Mnx and Irx transcription factors during early nephrogenesis and open a new door for a whole series of further gain and loss of function experiments to better understand how Sox32 modulated these targets for the normal tubule morphogenesis and proper nephron function.

Mutations of Sox32 also affects early heart development with defects in cardiac fusion (Alexander et al., 1999). This role of Sox32 was also recognisable in my transcriptomic analysis. At 9.00 hpf, Sox32 was involved in controlling a subset of genes which control cardiac morphogenesis, for example the zinc finger transcription factors Gata5 and Casz1. Gata5 is indispensable for zebrafish cardiac development and *gata5* mutant embryos lack a primitive heart tube and foregut (Holtzinger and Evans, 2007; Lou et al., 2011; Reiter et al., 1999; Wen et al., 2017). In *Xenopus*, loss of Casz1 results in cardiac defects from reduced myocardial integrity, improper deposition of basement membrane and a subsequent failure of cardiac cells to undergo cell movements associated with cardiac formation (Liu et al., 2014; Sojka et al., 2014). CASZ1 is also expressed in murine cardiomyocytes where it regulates cell cycle progression in both the first and second heart fields (Dorr et al., 2015). The observation that Sox32 regulates cardiac morphogenesis would explain why heart field fusion is impaired in *sox32* mutant embryos and complete cardia bifida often occurs (Dickmeis et al., 2001).

Of particular interest was the integration of the Sox32 downregulated gene list with the *sox17:GFP* enriched datasets at 9.00 hpf, which not only expanded the numbers of reciprocal genes that link endoderm-mesoderm interactions mediated by *sox32* and *sox17* but also allowed to highlight the similarity in the regulatory interactions between Sox32 module and Sox17 module during endoderm development.

I found that not only were the genes related to cardiac development (*gata5*, *casz1*) present in GFP^{+} cells, but also other well know markers of cardiac progenitor cells or differentiated cardiomyocytes including *tbx1, isl1*, *mef2cb,* and *tpm4b* expression of which were also confirmed by RT-qPCR.

Tbx1 is required for second heart field proliferation in zebrafish; *tbx1* (*van gogh, vgo*) mutants show an undersized ventricle, decreased number of cardiomyocytes and impaired migration of pharyngeal cells into the heart tube (Nevis et al., 2013). Similarly, Isl1 which is the an established marker of second heart field progenitor cells in mouse, was enriched in GFP^{+} cells. Islet family members are LIM homeobox transcription factors which are expressed both in cardiac progenitor cells and in pancreatic cells (Argenton et al., 1999; Dalgin et al., 2011; Wilfinger et al., 2013; Witzel et al., 2017; Witzel et al., 2012). *mef2cb* is important for the differentiation of both the first and second heart field cardiomyocytes (Hinits et al., 2012) and *mef2ca* is crucial for adding cardiomyocytes to the arterial pole of heart. Interestingly, mef*2ca* was exclusively downregulated in the *mixl1* dataset at 5.25 hpf (Hinits et al., 2012; Lazic and Scott, 2011). This highlights how *sox17:gfp* transcriptomic is partially the sum of target genes of Sox32 and Mixl1, which activate some shared and some divergent *cis*-regulatory module.

Additional target genes involved in cardiac morphogenesis were *tpm4a* and *tpm4b* which regulate embryonic heartbeat in zebrafish, *tpm4a* was found only in GFP^{+} cells whereas *tpm4b* isoform was downregulated in both the *mixl1* 5.00 hpf and *sox32* 9.00 hpf datasets (Dube et al., 2017; Wu et al., 2008). Other heart development related transcripts identified in my datasets were *atp2b1b* (*mixl1* 5.25 hpf) and s*lmapb (sox32* 9.00 hpf), both of which are expressed in the bulbus arteriosus (Singh et al., 2016). Lastly, Pbx genes are required in zebrafish early heart development and were observed to be less abundant in *mixl1*$^{-/-}$ compared to WT at 5.25 hpf (*pbx1b, pbx3b and pbx4*) (Maves et al., 2009). Overall, these observations on relatively different contributions of Mixl1, Sox32 and Sox17 to cardiac related genes expression substantiate and further expands the notion of diverging regulatory functions among TFs, with their turnover in regulatory motifs correlating with changes in regulatory activity and therefore how the temporal control of gene expression is integrated within a developmental network with a precise output (Naval-Sánchez et al., 2015; Potier et al., 2014).

The pancreas also derives from endodermal progenitors. Some target genes involved in this organ morphogenesis were previously mentioned to be downregulated in *sox32* mutant, such

as *jag1a* but data mining of the GFP^{+} dataset continued to expand the list of important pancreatic genes. Genes such as the alpha cell marker glucagon (*gcga*), regulators of pancreatic cell differentiation (*insm1b*) and insulin (*insb*) were validated to be enriched in the endodermal transcriptomic signature (Osipovich et al., 2014; Papasani et al., 2006; Tarifeño-Saldivia et al., 2017). These findings corroborate the importance of Sox32 in regulating pancreatic fate and the power of using FACS on *sox17:gfp* transgenic line to separate cell population marking multiple endodermal relevant trajectories and adding pancreatic fate to the previously described pharyngeal arch and heart formation.

My previous results demonstrate that both *mixl1* and *sox32* mutant embryos present defects in gene expression of multiple tissues and mesendodermal precursors cannot complete endodermal differentiation, however during development, not only cell identity through activation of specific cohorts of genes but also tissue morphogenesis must be finely orchestrated. Although many of the molecules that induce mesendoderm have been recognised, much less is known about the cellular mechanisms underlying mesendodermal cell internalisation and germ layer formation. To address this question, I then focus on identifying signalling molecules that might controls endodermal migration during zebrafish gastrulation.

As noted in Chapter 2, at the late blastula stage (4.00 hpf), endodermal and mesodermal progenitors are located in partially overlapping territories around the margin; with the start of gastrulation at around 5.00 hpf, segregation and migration of the cells of these 2 germ layers begins. The exact nature of these gastrulation movements however, remains unclear. Whereas a clearer understanding of the molecular pathway is taking shape and expression of specific markers that discern endodermal from mesodermal (Sox32 and Sox17) cells is being elucidated, the molecular mechanisms of morphogenesis that establish these boundaries are still not clear. 2 concurrent pathways are present in endoderm and mesoderm cells, one that specifies molecular identity and one that regulates movement. This aspect of endoderm development is highlighted by transplant experiments where, for example, both Nodal signalling and *sox32* expression need to occur simultaneously in order to have proper endoderm development (Liu et al., 2018; Rogers et al., 2017). Cell migration involves complex rearrangements of the actin cytoskeleton to position cells in their correct locations; this is coordinated by numerous remodellers and regulatory proteins (Montero et al., 2005; Schepis and Nelson, 2012) with cells responding to multiple dynamic migratory cues. Signalling and patterning information change intensity over time during different phases of migration. In

respect of this, I observed that several members of the laminin, cadherin, integrin, chemokine, cytokine and GTPase families were also affected in the analysed mutant embryos and specific signalling molecules were enriched in in GFP^+ cells. This supports the hypothesis that endodermal cells exhibit progressive changes in migratory behaviour and dynamics during gastrulation.

Starting with the Laminin family, a trio of *laminin* genes: *lama1* (laminin, alpha 1), *lamb1a* (laminin, beta 1a) and *lama5* (laminin, alpha 5) were all found to be significantly downregulated at both time points in both mutants and correlatingly more abundant in the GFP^+ transcriptome. All 3 laminins are heterotrimeric glycoproteins and have been shown to be involved in central nervous system development, defects in notochord differentiation and alterations in the retinas in zebrafish (Biehlmaier et al., 2007; Parsons et al., 2002). *lama5* is also crucial for formation of the apical ectodermal fold during fin development (Webb et al., 2007). These proteins act as cellular receptors for integrins and other cell surface molecules and are crucial elements of the extracellular matrix, indeed both cytoskeletal movement and receptor reorganization are dependent on laminins (Smyth et al., 1999). In mice, loss of LAMC1 results in embryonic lethality due to failure of endoderm differentiation (Smyth et al., 1999); however, the role of Laminins in zebrafish endoderm has not yet been studied.

Another group of interesting proteins that I found to be less abundant in both *mixl1* and *sox32* mutants compared to the WT embryo transcriptome were Cadherins, transmembrane proteins which are a type of cell adhesion molecule (CAM) important in the formation of adherence junctions which facilitate cell-cell interactions. Cadherins have wide-ranging roles during early embryogenesis, from regulating cell movements and tissue formation to brain development and neural crest cell migration (Babb et al., 2001; Clay and Halloran, 2014; Schepis and Nelson, 2012; Straub et al., 2011; Warga and Kane, 2007). I found several Cadherins undergoing dynamic changes in expression throughout my datasets with *cdh6* (*cadherin 6*) being statistically the most downregulated in *sox32* mutant. This is the first time that *cdh6* protein has been associated with endoderm development in zebrafish. Other predominant downregulated components were *cdh1*, *cdh2*, *cdh12a*, *cdh19* and *cdh23.* In *cdh1* morphants, gastrulation cell movements fail, due to defects in convergence and extension between mesodermal and endodermal cell layers and alteration in migrating cells toward the midline and animal pole (Babb and Marrs, 2004; Montero et al., 2005). Cadherin2 is essential for morphogenesis of the mesodermal germ layer during gastrulation and plays roles in the

formation and function of the heart, with the *cdh2* mutant showing an abnormally sized heart with an enlarged pericardial cavity and disorganized atrium and ventricle (Bagatto et al., 2006; Warga and Kane, 2007). Giger and David (2017) recently discovered that *cdh2* expression triggers endodermal cells to actively internalise at the margin of the embryo and migrate away from neighbouring cells (e.g. mesodermal) during gastrulation in a process mediated by Rac1, thus revealing cell contact avoidance as a previously unexplored mechanism driving endoderm formation. *rac1a* was significantly upregulated in *mixl1* dataset, suggesting that endodermal cell motility and actin dynamics via Rac1 and Prex1 could be drastically affected in this mutant. Signals that initiate and coordinate endodermal cells migration have never been studied in *mixl1* mutant, further experiments to characterise migration patterns are therefore needed.

Analysis of my data also revealed that Protocadherins play an important role during early endoderm development. Protocadherins, a subclass of the larger family of Cadherins, has also been shown to be involved in regulation of cell movements during gastrulation. My data revealed that *pcdh8, pcdh9, pcdh20, pcdh10b, pcdh2ab10, pcdh2g13, pcdh2ab6* and *pcdh2ac* were all downregulated in the analysed mutants and enriched in GFP^{+}. Notably, *protocadherin 8* (*pcdh8* or *papc*) is involved in morphogenesis of gastrula mesoderm and is a direct downstream target of *tbxta* and *tbx16* (Pei et al., 2007; Yamamoto et al., 1998), which are important mesendodermal TFs. In addition, *pcdh8* is downregulated by *sox32* overexpression and Fgf signaling can rescue its expression (Mizoguchi et al., 2006). The role of the other *protocadherins* in zebrafish embryonic development is poorly understood, and my data clearly shows how Sox32 and Mixl1 modulate the expression of multiple *pcdh* genes to ensure correct cell movement in the embryo and the correct segregation of the germ layer progenitors. It is possible that a deeper characterisation of this family of Cadherins would reveal further insights into the movement of neighbouring cells during endoderm internalisation.

Large scale cell movement during gastrulation sets up the body plan of the embryo and these rearrangements rely on different cellular mechanisms at different times and domains to physically create endodermal and mesodermal layers in the embryo, therefore a large repertoire of coordinated proteins is required. I questioned if Sox32 and Mixl1 were regulating this *cadherins* genes differently. I found that *snai1a* and *snai1b,* transcriptional repressors of E-cadherin expression (Montero et al., 2005; Yamashita et al., 2004) were significantly less abundant in the *mixl1* mutant but more abundant in the *sox32* mutant. The most likely

explanation for this divergence is that both the Sox32 and Mixl1 modules regulate the expression of multiple E-cadherins, but their subsequent downregulation by *snail* genes creates a different readout of cell-cell adhesion and therefore a different adhesion-dependent morphogenesis. This in turn creates different states of cellular motility between cell types and could also potentially explain the separation of mesendodermal cells. This speculation is supported by findings in *Drosophila,* where Snail and Twist interact with Sp1 to separate the endoderm from the mesoderm prior to gastrulation (Bronner et al., 1994). In addition (Qiao et al., 2014) observed that *snail* genes control the morphogenesis of the heart in zebrafish embryos by modulating the extracellular assembly of fibronectin via the expression of α5 integrin; *snail* morphants display disrupted migration of cardiac precursors. 2 noticeable observations are therefore that: i) Snails are zinc finger TFs like *gata5* and *casz1* as previously described and ii) my datasets are enriched for integrins *(itga3a, itga4, itga5, itga8, itga10* and *itga11a).* It has been shown that cardiac precursors use endodermal cells as physical substrate to migrate (David and Rosa, 2001; Lough and Sugi, 2000). The dysregulation of integrins and cardiac genes expression observed in my data is a strong candidate for the molecular mechanisms leading to the observed cardia bifida phenotype in *mixl1* and *sox32* mutants. Integrins mediate the link between actin stress fibres of the cytoskeleton and the extracellular matrix and this connection provides the traction forces for migration (Huttenlocher and Horwitz), and thus alteration of this platform can explain the failure of myocardial migration in *sox32* and *mixl1* mutants.

In mouse and *Xenopus,* multiple studies have linked downstream targets of Nodal signalling (Mix-like factors, Eomes, Lim1, Foxa2, and Gata4–6) to mesendodermal migratory behaviour (Arnold et al. 2008; Fletcher et al. 2006; Kofron et al. 2004; Luu et al. 2008; Tam et al. 2004, 2007). In zebrafish, Nodal induced TFs together with the Mixl1 paralogue, Sebox, activate the expression of chemokine receptor Cxcr4 (Dickinson et al. 2006, Fukui et al. 2007, Sinner et al. 2006) and the ligand Cxcl12b, the latter of which acts as a chemoattractant for *cxcr4* expressing endoderm cells (Fukui et al. 2007, Mizoguchi et al. 2008, Nair & Schilling 2008). The disruption of *cxcr4*/*cxcl12b* in zebrafish results in disrupted endoderm migration and gut-tube duplications (Mizoguchi et al. 2008, Nair & Schilling 2008). *cxcr4*/*cxcl12b*, as well as PDGF signalling, appear to act by regulating integrin fibronectin-mediated endoderm migration (Keller 2005). Similar to the *snail* genes, Mixl1 and Sox32 regulate Cxcr4a oppositely in the mutants with *cxcr4a* being upregulated in the former and downregulated in the latter. Chemokines are small secreted proteins implicated in cell migration in various

biological processes and belong to the superfamily of G-protein-coupled receptors (Kucia et al., 2004, Busillo and Benovic, 2007). Consistent with a large body of evidence in the literature, *cxcr4*/*cxcl12b* signalling regulates ECM-integrin-dependent adhesion during gastrulation cell movements (Nair and Schilling, 2008) with *cxcl12b* expressing mesodermal cells acting as a chemoattractant and directionally controlling the movements of *cxcr4a* expressing endodermal cells.

Speculatively, my data extend this information to include opposite regulatory functions of Sox32 and Mixl1. Sox32 activates *cxcr4a* and represses *cxcl12b* in WT embryos and Mixl1 represses *cxcr4a* and activates *cxcl12b.* This interaction could be another, additional, regulatory motif during endoderm development that helps in channelling endodermal cells with different fates. Mizoguchi et al. (2008) speculated that *cxcr4a* expressing endodermal cells are guided by the overlying *cxcl12b* mesodermal cells to the dorsal side of the embryo during gastrulation. Integrating this theory into my data, *sox32* expressing cells would migrate earlier or with a different orientation than *mixl1* expressing cells, thus giving these cells different directionalities and/or time windows in which to migrate. In WT embryos, inhibition of *cxcr4a* or *cxcl12b* delays endodermal migration by reducing the number of filopodia in endodermal cells and in the *sox32* mutant, the *cxcr4a* 'salt and pepper' expression pattern is absent whereas expression of *cxcl12a* and *cxcl12b* is unchanged. Further experiments to detail the spatial expression dynamics of *cxcr4a, cxcl12a* and *cxcl12b* in the *mixl1* mutant are required, however taken together, these data suggest that dissimilar or coordinated cellular recruitment of chemokine may drive relevant specific movement in endodermal cells subsets.

Recently*,* (Collins et al., 2018) provided new insight into the cohort of endodermal genes that promote actin dynamics and migration, with the discovery of *pitx2c* and the fibronectin receptor subunit gene *itgb1b (integrin, beta 1b). pitx2c* was not found to be differentially expressed in my Sox32 or Mixl1 datasets, however I did find *itgbl1* (*integrin, beta-like 1*). RT-qPCR analysis validated the downregulation of this gene in the *mixl1* mutant, hence I can hypothesise that similar to *itgb1b, itgbl1* cooperates in a network of integrins to drive mesendodermal cell migration during gastrulation.

Strikingly, fibronectin-leucine rich transmembrane protein (Flrt3) was a common component in all datasets, and Flrt2 was found to be downregulated in the *mixl1*$^{-/-}$ 5.25 hpf and enriched in the GFP^{+} datasets. Defects in migration of definitive endoderm are observed in FLRT3 null mouse embryos (Egea et al. 2008, Maretto et al. 2008) and Ogata et al. (2007)

showed that FLRT3 controls cadherin-dependent cell adhesion and mesendodermal migration via the small GTPase RAND1 (Ogata et al. 2007). Flrt3 could therefore be another novel potential player involved in morphological movement in zebrafish during gastrulation.

The Rho family of small GTPases are another class of established regulators of cell migration, for example, RhoA, Rac1 and Cdc42 proteins transduce signals, influence cell behaviours and play several well studied roles in regulating actin dynamics during cell migration (Ridley et al., 1992; Woo et al., 2012) evidenced that Nodal signalling can affect actin stability and migration in endodermal cells and they were able to link these changes to the action of Rac1, and the expression of the Rac activator Prex1. The latter was significantly downregulated in *sox32*$^{-/-}$ and considerably enriched in GFP^{+} at 9.00 hpf. Concurrently, I also found in my analysis Rasl11b, multiple Cdc42 - guanine nucleotide exchange factors (GEFs) (*arhgef5, arhgef37, arhgef1a, arhgef7b*) and multiple Rho GTPase activating proteins (*arhgap5, arhgap17b, arhgap21b, arhgap23, arhgap24, arhgap29b, arhgap32b, arhgap33, arhgap35a, arhgap35b*). All of these genes have uncharacterised roles in zebrafish. Cdc42 effector protein has been studied in the context of molecular mechanisms controlling polarization (Etienne-Manneville, 2004) and can directly stipulate nucleation of actin filaments via its effect promoting factors including WASP and the Arp2/3 complex (Yang et al., 2000). Future studies should aim to better characterise the role of these small Rho GTPases, knowing that i) endodermal cells show characteristic filopodia when migrating (Mizoguchi et al., 2008) and ii) Rac1 and Arp2/3 are required for the internalization of endodermal cells (Giger and David, 2017).

An atypical cytoplasmic Ras small GTPase**,** *rasl11b* was also found to be downregulated by Sox32 at 9.00 hpf. Knock down experiments have previously revealed *rasl11b* as a negative modulator of endoderm and prechordal plate formation and demonstrated its ability to partially rescue zygotic *tdgf1*$^{-/-}$ mutants. This supports the theory that mesendoderm formation is orchestrated by 2 parallel pathways: Nodal type I dependant receptors and Nodal type II dependant receptors with the Tdgf1 factor playing a role in parallel to, or upstream of, the Nodal ligand/receptor complex. The constitutively active Nodal type I receptor *acvr1ba* is able to commit cells to an endodermal fate and rescue the MZ*tdgf1* mutant phenotype. The *tdgf1* gene is possibly necessary as a coreceptor for Nodal signal transduction via the serine/threonine kinase receptor complex and selectively targets type I and not type II receptors in the complex. In this way, a similar role for *rasl11b* should be investigated.

prex1, a Rac-GEF that I previously mentioned discussing the role of GTPases*,* is not only a direct target of Nodal signalling, promoting Rac1 activity and regulating endodermal cell motility – but is also under the control of Sox32. Further experiments to validate the potential role of Rasl11b and Cdc42 are needed not only in *sox32* mutants but also in *mixl1* mutants; comparing the expression in the different mutants should help to establish their role(s) in modulating the dynamic motility of endodermal cells, as well as their internalisation in more detail.

Recent work has begun to reveal the molecular mechanisms that link endoderm formation and patterning with the cell migration, cell adhesion and cytoskeletal dynamics that control endoderm morphogenesis. My data are an additional, valuable asset to identify novel genes important to endodermal migration at the gastrula stage. It is also very likely that other cytoskeletal regulatory proteins besides the ones I have described are involved in endoderm morphogenesis. Indeed, in my transcriptomic analyses, I identified several genes associated with cell migration and cytoskeletal dynamics. Future studies and further mining of the datasets will likely identify additional cytoskeletal regulators important for tissue morphogenesis and organ development.

In conclusion, my analysis has corroborated the existing knowledge of gene activity for *mixl1* and *sox32* during zebrafish gastrulation, but in addition, has also revealed novel participants in endoderm development. My results identified new target genes of Mixl1, recognised the potential role of non-coding RNAs and highlighted Sox32 as a TF with an amazing capacity to coordinate and control multiple signal transduction pathways.

Gastrulation movements occur within a dynamic environment and I have illustrated how Sox32 and Mixl1, by controlling different targets or sharing them, could actively and precisely control spatiotemporal cell movements, and that this coordination could be achieved using a wide variety of extracellular cues during gastrulation. The uniquely and differentially expressed cohorts of laminins, cadherins, integrins, chemokines, cytokines and GTPases will provide valuable information to help us further understand how endodermal cells undergo developmentally regulated changes in migratory behaviour.

In conclusion, in this chapter I have presented new avenues of investigation into the endodermal development in zebrafish embryos. I have shown how different TFs can orchestrate different aspects of endoderm development through regulating, in parallel, a

network of other TFs and signalling pathways. The findings presented in this chapter have clear implications in further characterising the zebrafish endodermal GRN. This comprehensive transcriptome can function as a reference catalogue for interpreting gene expression while modelling endoderm cell fate decisions in zebrafish. As the next step in my project and as detailed in the next chapter, I proceeded to organise all the information I had collated and integrated, and, using methods for inferring transcriptional regulatory networks from gene expression profiling data, began to build a new endodermal GRN.

Chapter 6 – Endodermal GRN during zebrafish gastrulation

Chapter 6 highlights:

- Description of multiple sources of biological data that are used for GRN analysis and inference.
- Integration of time series analyses with transcriptomic analysis of perturbed endodermal systems.
- Integration of TFs information.
- Further exploration of the zebrafish endoderm regulatory network.

6.1 Introduction

As detailed in Chapter 1, a major challenge in biology today is to understand the processes that control formation of complex organisms. During development, a single cell, the fertilised egg (zygote), is transformed into a mature adult comprised of millions of cells and dozens of organs, each with unique identity and function. To understand this complex process, we first have to look into how genes are activated and how different cell types are specified during embryonic development. In 2009, Chan et al., published the first GRN underpinning zebrafish development, in which they attempted to describe in detail how the different genes and morphogens involved in early development are interconnected. What is particularly notable from this work, is the lack of information regarding endoderm development, compared with that of ectoderm and mesoderm. Since 2009, further studies have been published in respect of endoderm development, however the focus has been primarily on early Nodal signalling dynamics involved in the bifurcation of mesendoderm (Rogers et al., 2017; Liu et al., 2018; van Boxtel et al., 2018; Vopalensky et al., 2018) and/or late somitogenesis and the specification of hepatopancreatic cells (Tarifeño-Saldivia et al., 2017; Zhang et al., 2017). None of these studies have attempted to characterise, in detail, the GRN underpinning endodermal specification (and cellular migration during this process), more specifically, the roles of zebrafish endodermal development regulators, Sox32 and Sox17.

More recently, in 2017, Nelson et al., published a paper in which they detailed the mesendodermal network using a combination of ChIP-seq, RT-qPCR and *in situ* hybridizations (Figure 6.1)

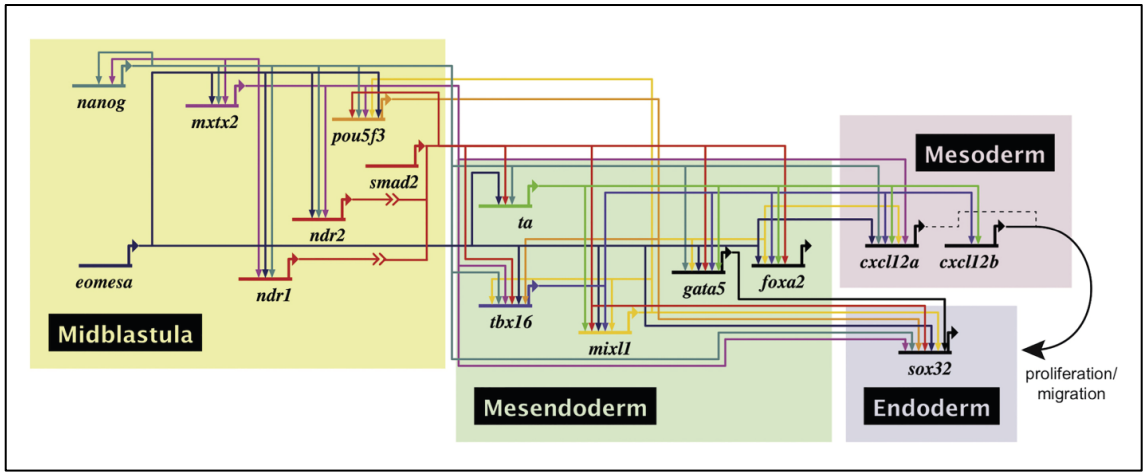


Figure 6.1 Mesendodermal GRN from Nelson et al., 2017. A GRN for endoderm formation informed by this study. Evidence of interactions was inferred by ChIP-seq analysis for multiple TFs combined with expression data collected from the literature. The authors identify three temporal/spatial domains illustrated by shaded boxes. At midblastula stage (yellow), a combination of maternal and non-maternal factors induce a set of TFs that create a transient cell population called mesendoderm (green). Endodermal fate specification proceeds (purple) through the combinatorial interactions of the abovementioned mesendodermal TFs which ensure the activation of *sox32* expression, the master regulator of zebrafish endoderm formation. Both *cxcl12a* (dotted line – marginal role) and *cxcl12b* promote endodermal cell proliferation and migration from the mesodermal domain (pink). ≫ indicates binding to Nodal ligand-receptor and intracellular Smad2 activation.

Although this study added more players to the GRN and established new connections, the gap in our knowledge regarding the endodermal GRN still persisted. The GRN preceding early endodermal induction had become clearer, but gene regulatory dynamics during the 5.25 hpf and 9.00 hpf window are still poorly understood. This is illustrated in Figure 6.1, where the endodermal GRN (purple, bottom right) simply shows *sox32* as an important TF but reveals no further information about its regulatory function. I therefore aimed to combine the existing data on endodermal development present in the literature with my own experimentally generated data to construct a more comprehensive endodermal GRN and bridge this knowledge gap.

In order to build an updated version of the network, I decided to combine four different approaches. The first step involved mining the existing literature, in particular, the information

on ZFIN, the online zebrafish database. Thereafter, I proceeded to gather data on gene expression dynamics during endoderm development using available genome wide RNA-seq time series datasets. I then combined these data with my experimental transcriptomic data in which the system was perturbated (*sox32*$^{-/-}$ and *mixl1*$^{-/-}$ mutants) and I added the information I generated from profiling only endodermal GFP^{+} cells isolated from the *sox17:GFP* transgenic line. Finally, I assessed the direct/indirect nature of the new identified connections using protein-DNA interaction data (ChIP-seq datasets). Below, I detail each approach individually and the criteria used to build network connections. If a GRN is constructed using a large amount of data collected on both spatial and temporal gene expression, it can tell us much about how different cells in the embryo are specified, and predictive models can be built with enough data to simulate cellular behaviour and predict the outcome of perturbation experiments (Linde et al., 2015; Chen et al., 2016; Cholley et al., 2018). Differences observed between experimental data and these simulations can consequentially help us to discern gaps in the regulatory processes of the currently known network architecture and provide the starting point from which to formulate new hypotheses to better describe the observations, which can then be tested accordingly.

Using a combination of these strategies, the Davidson and Peter labs have reconstructed the GRN that controls the development of mesendoderm in sea urchin; this GRN comprises more than 100 regulatory and signalling genes and resolves multiple nodes along the specification of mesendodermal cells. In addition, they introduce a computational approach (Boolean model) to help describe the spatial and temporal gene expression and gene interactions during sea urchin gastrulation (Oliveri and Davidson, 2004; Davidson, 2009; Peter and Davidson, 2010; Erkenbrack et al., 2018). They also incorporate new parameters into the model including embryonic geometry and gene expression kinetics to help predict and explain gene expression patterns (Bolouri and Davidson, 2003).

Since then, other groups in the zebrafish community have started exploiting the power of GRNs and have begun to model these diverse networks. For example, Greenhill et al. (2011) combined experimental observations with mathematical modelling to explore the core melanocyte GRN, Petratou et al. (2018) used similar methods to study iridophore specification, and other studies have generated the GRN underlying neural crest development (Simões-Costa and Bronner, 2015; Martik and Bronner, 2017; Williams et al., 2018).

6.2 Data mining

What do we already know about the genes involved in endoderm formation? As detailed in Chapter 1, recent studies, have begun to elucidate the developmental mechanisms that control the induction and patterning of the endoderm in different model organisms. Comparison of three common models, mouse, *Xenopus* and zebrafish, showed that the endodermal gene pathway is generally conserved - the same families of TFs and signalling pathway are involved, including Nodal family members, Mix, Gata and Sox TFs. In respect to the Sox family, the situation in zebrafish is more complex due to genome duplication – Sox17 (Figure 6.2, shown in green) is the master regulator of endodermal fate in frog and mouse (Hudson et al., 1997; Kanai-Azuma et al., 2002; Sinner et al., 2004; Niakan et al., 2010), whereas in zebrafish this task is performed by the closely related Sox32. Sox32 is unique to zebrafish and appears to be a crucial regulator of endodermal versus mesodermal fate (Poulain et al., 2006) and the regulation of its activity is likely to be essential for the proper ratio between endoderm and mesoderm cells. Embryos deficient in *sox32* lack all endoderm structures and develop cardia bifida (Alexander et al., 1999; Dickmeis et al., 2001; Kikuchi et al., 2001).

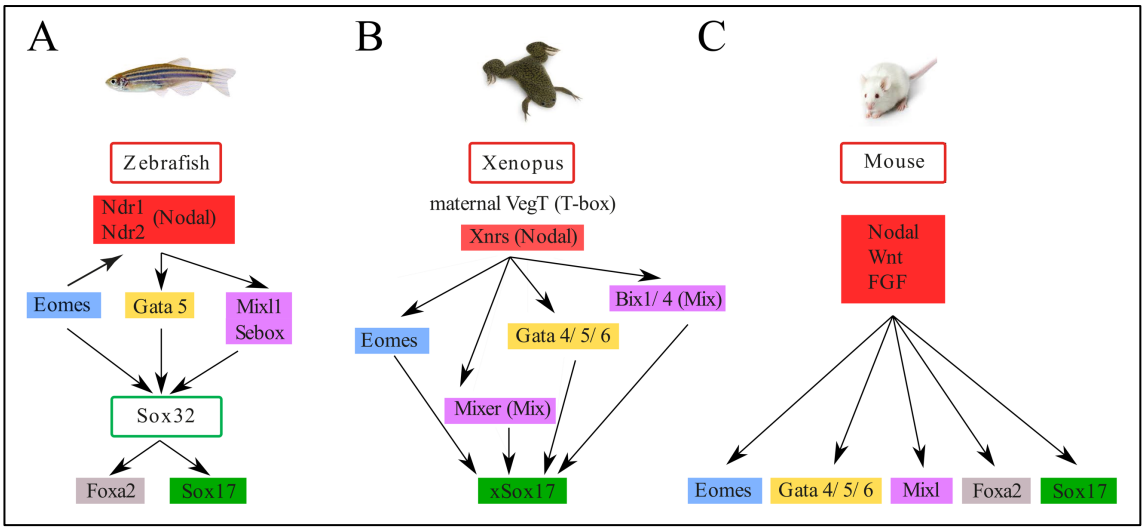


Figure 6.2 Conserved pathway depicting endodermal specification in zebrafish (A), *Xenopus* (B) and mouse (C). TFs families are colour coded. Note that i) Sox32 is only present in zebrafish and ii) the hierarchical structure of the interactions changes between the species. The functions of the downstream effectors of Nodal are highly intertwined and they can regulate endoderm specification, but the factors are positioned at different levels of the signaling cascade depending on the species. For example, Eomes is directly under the control of Nodal in *Xenopus* and mouse but in zebrafish Eomes regulates Nodal signal.

The first step I took to obtain a systematic view of the coordinated activity of multiple TFs in endoderm development was to search the literature, annotate the interactions between these

factors and summarise the information in a transcriptional GRN. I collected the present knowledge starting from the induction of mesendodermal precursors at around 3.25 hpf through to the end of gastrulation at 10.00 hpf when patterning and differentiation starts. Most of the data were limited to traditional approaches including *in situ* hybridization, RT-qPCR and gene knockdown/knockout. Importantly, these data, despite being readily available, were not yet systematically integrated into a GRN model.

In addition to scanning the literature, I further collected and assembled the information available on the Zebrafish Model Information Network (ZFIN) (Ruzicka et al., 2015) which is a curated database that collects zebrafish papers and reports both spatial and temporal expression patterns of annotated genes. The availability of databases with central information on genes is an essential resource when building a GRN. Databases such as FlyBase (FlyBase et al., 2018), XenBase (James-Zorn et al., 2015) and EchinoBase (Kudtarkar and Cameron, 2017) have been indispensable for dissecting the GRNs of flies, *Xenopus* and sea urchin respectively. Multiple resources are needed to generate a GRN including visualisation tools, literature evidence, TF binding databases and gene expression databases.

Similar to the approach described by Chan et al., (2009), I downloaded the following text data files from the ZFIN website: 'expression data for wild type fish.tsv', 'zebrafish stage series.tsv' and 'zebrafish gene expression by stage and anatomy term.tsv'. These data files contained all the available data on mRNA *in situ* hybridizations and temporal expression domains of all annotated genes. I then proceeded to summarise this information in a coherent way to better visualise it with regards to the period of development as well as localisation within the embryo. At this stage, I limited the information for each gene to expression in wild type fish. The correlation of both spatial and temporal expression domains amongst genes was important to draw conclusions about the input and output of the GRN. A prerequisite for the network was that genes needed to be coexpressed in a spatiotemporal manner in order to infer regulation.

I found 144 genes that were associated with 'hypoblast', 'mesendoderm', 'presumptive endoderm', 'endodermal cell' and/or endoderm queries to construct a global set of gene expression. These data were then reorganised into a table that reported at what developmental stage, and in which tissue the genes were expressed, similar to the table used in the generation of the GRN underlying sea urchin mesoderm formation and the most recent zebrafish GRN (Chan et al., 2009b; Peter and Davidson, 2011b). An example of the resulting tables is reported

in Figures 6.3 and 6.4. To achieve this result, I extracted ZFIN's descriptions of anatomical terms reporting the anatomical structures they belong to, as well as known substructures. With this information, I built hierarchies of anatomical systems and for each structure assigned specificity. Some entry terms did not have defined structures, and I had to manually curate those and define whether they were related to endoderm or not.

A

	Zebrafish Stage Series	Hpf	Expression range: Sox32	Sox17	Gata5	Pou5f3
blastomere	Blastula:512-cell	2.75				x
	Blastula:1k-cell	3				x
	Blastula:High	3.33				x
	Blastula:Oblong	3.66				x
presumptive structure	Blastula:Sphere	4	x			x
	Blastula:Dome	4.33	x			x
	Blastula:30%-epiboly	4.66	x	x		x
	Gastrula:50%-epiboly	5.25	x	x	x	x
	Gastrula:Germ-ring	5.66	x	x	x	x
	Gastrula:Shield	6	x	x	x	x
primary germ layer	Gastrula:75%-epiboly	8	x	x	x	x
	Gastrula:90%-epiboly	9	x	x	x	x
	Gastrula:Bud	10	x	x	x	x
	Segmentation:1-4 somites	10.3	x	x	x	x
	Segmentation:5-9 somites	11.7	x	x	x	
	Segmentation:10-13 somites	14	x		x	
	Segmentation:14-19 somites	16	x		x	
	Segmentation:20-25 somites	19			x	
	Segmentation:26+ somites	22			x	
	Pharyngula:Prim-5	24			x	

B

	Zebrafish Stage Series	Hpf	Spatial domain for Sox32: YSL	presumptive endoderm	endoderm cell	margin	hypoblast	unspecified
blastomere	Blastula:512-cell	2.75						
	Blastula:1k-cell	3						
	Blastula:High	3.33						
	Blastula:Oblong	3.66						
presumptive structure	Blastula:Sphere	4	x					
	Blastula:Dome	4.33	x					
	Blastula:30%-epiboly	4.66	x	x		x		x
	Gastrula:50%-epiboly	5.25	x	x				x
	Gastrula:Germ-ring	5.66	x	x				x
	Gastrula:Shield	6	x	x				x
primary germ layer	Gastrula:75%-epiboly	8	x		x		x	x
	Gastrula:90%-epiboly	9	x		x			x
	Gastrula:Bud	10	x		x			x
	Segmentation:1-4 somite	10.3	x		x			x
	Segmentation:5-9 somite	11.7			x			
	Segmentation:10-13 somi	14			x			
	Segmentation:14-19 somi	16						
	Segmentation:20-25 somi	19						
	Segmentation:26+ somite	22						
	Pharyngula:Prim-5	24						

Figure 6.3 Expression table summarising both temporal (left) and spatial (right) expression of endodermal associated genes as labelled. To build the GRN I downloaded and organised the information from the log files of ZFIN. Top panel (**A**) shows a time series of gene expression for *sox32, sox17, gata5* and *pou5f3*. These 4 TFs are coexpressed between 5.25 hpf and 10.33 hpf. In the bottom panel (**B**), spatial expression domains for *sox32* inferred from WISH data are reported.

A

structure	stage	gene	<16	16-22	>22
endoderm	22	aldh1a2	-	+	-
endoderm	24	ass1	-	-	+
endoderm	26	ass1	-	-	+
endoderm	25	bmp2b	-	-	+
endoderm	26	bmp2b	-	-	+
endoderm	27	bmp2b	-	-	+
endoderm	28	bmp2b	-	-	+
endoderm	29	bmp2b	-	-	+
endoderm	20	bmp7a	-	+	-
endoderm	29	bmp7a	-	-	+
endoderm	26	cldn15a	-	-	+
endoderm	27	cldn15a	-	-	+

B

pub	gene	unique gene	unique pub	sort by pub	gene	sort by gene	pub
imputed	cldn15a	ackr3b	imputed	ZDB-PUB-021105-10	sox17	sox32	ZDB-PUB-010810-1
imputed	cldna	agr2	ZDB-PUB-000309-43		sox32		ZDB-PUB-010810-1
imputed	cldna	aldh1a2	ZDB-PUB-000616-3				ZDB-PUB-010810-1
imputed	cldna	ass1	ZDB-PUB-001017-2				ZDB-PUB-010810-1
imputed	cldnb	bambia	ZDB-PUB-010219-1				ZDB-PUB-010810-1
imputed	foxa3	bmp2b	ZDB-PUB-010810-1				ZDB-PUB-010810-1
imputed	lcp1	bmp7a	ZDB-PUB-011214-14				ZDB-PUB-021105-10
imputed	nr2f1a	bmper	ZDB-PUB-021105-10				ZDB-PUB-040109-3
imputed	nr2f1a	cdx4	ZDB-PUB-021106-13				ZDB-PUB-061010-11
imputed	sox17	cldn15a	ZDB-PUB-030307-2				ZDB-PUB-070122-23
imputed	sox17	cldna	ZDB-PUB-031103-23				ZDB-PUB-071125-22
imputed	sox17	cldnb	ZDB-PUB-031103-24				ZDB-PUB-080226-5
ZDB-PUB-000309-43	bmp7a	col2a1a	ZDB-PUB-031103-3				ZDB-PUB-080414-12
ZDB-PUB-000309-43	bmp7a	ctgfa	ZDB-PUB-040109-3				ZDB-PUB-080414-12

Figure 6.4 Examples of additional tables used to visualise the spatial and temporal expression of endodermal genes. (A) Genes were associated with 3 ZFIN standardised developmental stages. < 16 is blastula stage, 16 to 22 gastrula stage and > 22 is associated with segmentation. Organising the data in this format meant that genes expressed at the same time point were easily identified. (**B**) Data were also organised by publication, to allow for easy identification of the reference that reported on a specific gene.

Although the data obtained from ZFIN encompasses multiple genes and can be organised by both time points and spatial domain, this information was still too fragmented and not always accurate. Many gaps in developmental time were detectable throughout the dataset, thus achieving sufficient temporal depth of resolution was not possible, and smooth expression trajectories of specific genes could not be determined. ZFIN data focuses primarily on individual genes and often does not shed any light at all on gene dynamics, as they are annotated as single points.

To circumvent these problems, I started interrogating existing genome-wide datasets that characterise transcript dynamics in the developing zebrafish embryo. Mathavan et al. (2005) were the first group to use microarrays to analyse temporal transcriptional events in zebrafish embryos, collecting data across twelve time points from the unfertilised egg to two days post fertilisation. Later, Yang et al. (2013) generated datasets for 9 different developmental time points covering developmental periods, from 2.0-2.2 hpf to 72 hpf. Other transcriptomic works

focused on specific developmental stages such as the maternal-zygotic transition (Aanes et al., 2011; Harvey et al., 2013; Lee et al., 2013), mapping specific long noncoding transcript (Pauli et al., 2012) or specifically studying the transcription start sites of important genes at high-resolution (Gehrig et al., 2009; Nepal et al., 2013). The most recent transcriptome baseline was generated by White et al. (2017) with an mRNA-seq expression time course across 18 time points during zebrafish development, from one cell stage to 5 days post fertilisation. This study characterised the temporal expression profiles of 23,642 genes. The combination of these datasets provided me with a starting point to study gene expression dynamics during the development of zebrafish embryos; the information that can be extrapolated from these data was: i) genes that behave in a similar manner, and ii) whether we can observe patterns of genes switching on/off. The aforementioned papers have answered some of these questions by revealing enrichment for stage specific biological pathways, and with clustering analysis were able to track progression of gene cohorts changing over time. Cohorts of significantly differentially expressed genes were matched, between stages with distinct expression patterns. The main purpose of these studies was to obtain a global view of the whole biological systems behaviour by using high-throughput approaches to extrapolate general expression patterns of all genes during development, then derive temporal synexpression to characterise genes of unknown function or to uncover new temporal phenomena in gene expression. However, due to aforementioned global approach, the specific endodermal gene expression pattern was not been investigated in these datasets. Further advanced analyses can be conducted on these datasets, with the proper bioinformatics methodologies, that can account for the intrinsic complexity and noise; for more information please refer to the following papers (Owens et al., 2016; Li and Li, 2018; Svensson et al., 2018). In this instance, a more basic approach was taken whereby the genes of interest were plotted and their temporal expression patterns observed as shown in Figure 6.7.

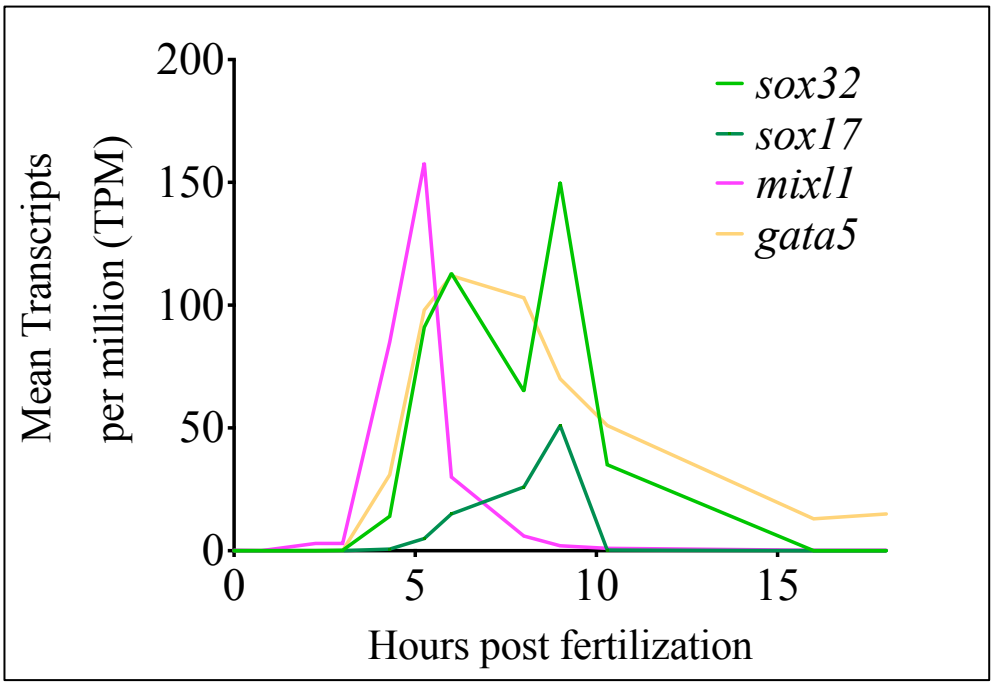


Figure 6.7 Temporal expression of important mesendodermal and endodermal genes. *mixl1* and s*ox32* expression levels peak at around 10 hpf. *sox17* levels are relatively low compared to the level of *sox32* and *mixl1*. *gata5* is expressed between 5 and 12 hpf. Expression profiles are shown as average expression (mean TPM) and were constructed from time series data (White et al., 2017).

The combination of the information obtained from ZFIN together with the high resolution transcriptional profiling of zebrafish development from the aforementioned studies was extremely powerful, as spatial and temporal gene expression were integrated together. A key limitation to this wealth of data however was that it was limited to the WT condition. The next step, in order to delineate new interactions in the endodermal GRN, was to integrate results where the system had been perturbated. This helps correlate changes in regulatory gene expression to changes in expression of target genes. Changes in regulation can be measured following gain-of-function experiments (e.g., injection of RNA encoding a TF) and/or loss-of-function experiments (e.g. injection of antisense morpholino oligonucleotides) and then examining correlated changes in RNA expression (*in situ* hybridization, RT-qPCR and RNA-seq). A representation of how these data can be summarised is reported in Table 6.1. This approach can be used to introduce gain and/or loss of function mutations in every step of a signalling cascade; changes in downstream effectors can then be observed, allowing one to discern the exact role(s) of the manipulated gene/signal.

Table 6.1 Summary of mutant line and morphants with endodermal defects. Expression pattern observed by whole embryo *in situ* hybridization or cross section are described.

Gene	Also known as	Nature	Phenotype
ndr1	*squint (sqt)*	TGF-β ligand	*ndr1* mutant: no prechordal plate and dorsal mesoderm defects. Mesendoderm development is only delayed
ndr2	*cyclops (cyc)*	TGF-β ligand	*ndr2* mutant: lack ventral midline cell types in the central nervous system
			ndr1/2 double mutant: all of the endoderm and most of the mesoderm do not develop, in addition anterior trunk spinal cord is absent
lft1	*lefty/antivin*	TGF-β ligand	Overexpression leads to absence of endoderm and dorsal mesoderm.
lft2	*Lefty/antivin*	TGF-β ligand	Overexpression leads to absence of endoderm and dorsal mesoderm.
acvr1ba	*taram-A (tar)*	Type I TGF-β receptor	Overexpression is able to commit cells to an endodermal fate by promoting *sox17* expression in wild type and *tdgf1* mutant embryos, but not in *sox32* mutants.
tdgf1	*one-eyed pinhead (oep)*	EGF-CFC coreceptor	Zygotic *tdgf1* mutat has no prechordal plate and endoderm. Maternal *tdgf1* mutant shows no prechordal plate, no endoderm and dorsal mesoderm.
gdf3	*vg1*	TGF-β cofactor	Gdf3 is required for mesoderm, endoderm and neural patterning. Morphants: L-R patterning defects. Zygotic *gdf3* mutants are viable and fertile. Maternal *gdf3* mutants have no notochord, spinal cord and structures associated with mesendoderm formation, endodermal tissues **and loss of gene expression domains marking axial mesoderm and lateral plate mesoderm**
nanog		Homeobox transcription factor	Maternal *nanog* mutant exhibits defects in epiboly morphogenetic movement, lack of axes formation and high percentage of cell death at the end of gastrulation

mxtx2		Homeobox transcription factor	Knockdown of *mxtx2* leads to a yolk burst phenotype similar to that observed in the *nanog-like* morphant
smad2		Substrate for the TGF-β family of receptors	Endoderm and head and trunk mesoderm are absent in maternal and zygotic *smad2* mutant, a phenotype very similar or identical to Nodal loss-of-function mutants
eomes		T-box transcription factor	Maternal zygotic *eomesa* mutants show a delay in doming and YCL microtubules defects
pou5f3	*oct4, spiel-ohne-grenzen (spg)*	Homeodomain transcription factor	Maternal zygotic mutant lack coordination of microtubules of the YCL and radial intercalation of deep cells thus epiboly is not complete. The mutants also show multiple patterning defects and increased apoptosis at the end of gastrulation
dusp4		Dual specific phosphatase	*dusp4* morphant displays loss of foregut and pancreatic endoderm
foxH1	*schmalspur (sur), fast1*	Winged helix transcription factor	Zygotic mutant: reduction of prechordal plate. Maternal zygotic mutant: no prechordal plate and reduced number of cells expressing endodermal markers during gastrulation
mixl1	*bonnie and clyde (bon) , mixer*	Homeodomain transcription factor	*mixl1* mutant shows 60% reduction of endodermal cells number, reduction in number of prechordal plate progenitor cells and cardia bifida.
sebox	*mix*, *mezzo*, *og9x*	Homeodomain transcription factor	*sebox* morphant embryo develops without any apparent defect. Overexpression of *sebox* mRNA induce ectopic expression of *sox32*, *sox17* and *tbxta*. *sebox* mRNA can rescue *mixl1* mutants, and *sebox* morpholino increases *mixl1* phenotype: no prechordal plate mesoderm and endodermal progenitors
gata5	*faust (fau)*	Zinc finger transcription factor	*gata5* mutant shows reduction of endodermal cells number (10%) with lower levels of *sox17* and *foxa2* expression. Embryos also present cardia bifida defect.
sox32	*casanova (cas)*	Sox Transcription Factor	*sox32* mutant and morphant: lack endoderm structures and develop cardia bifida. Brain defects are also visible
sox17		Sox Transcription Factor	*sox17* morphant shows defects in forerunner cell group morphology, Kupffer's vesicle formation and determination of left/right symmetry, abnormal pancreatic development, blood circulation is disrupted and pericardial oedema is visible

foxa2	*axial*, HNF3β	Winged helix transcription factor	Double knockdown of *foxa2* and *foxa3* prevent the formation of all axial derivatives while over-expression of increases dorsal mesodermal domain.

The benefits and limitations of using mutants and morphants to study the consequences of a genetic loss of function are well documented in zebrafish literature, for more details see Bedell et al. (201), Vogan (2015) and Stainier et al. (2017). Generally speaking, observation of phenotypical defects in the mutants/morphants compared to the WT is a fast way to discriminate gene function and associate phenotypic defects with the specific germ layer(s): endoderm, mesoderm or ectoderm. In some cases, often due to functional redundancy as a consequence of gene duplications, no defect is observed in a knockout zebrafish model (Rossi et al., 2015). This redundancy thus renders the developmental process resilient to perturbations. A robust GRN will generate constant biological output even in the presence of a perturbation, however the structure of the network can still change to accommodate compensatory mechanisms – even if the outcome is identical to the non-perturbed system.

The difference between gene expression patterns (network output) of mutated vs. WT networks can be quantified and added as another layer of information to the GRN. Transcriptomic analysis scales up this concept, quantitatively assessing changes in transcript levels of the whole population. When I compared gene expression levels in endodermal mutant fish and WT fish, I expected that the up or downregulated genes changes would be associated with the mutation profile. Nonetheless, as powerful as this method can be, the most limiting aspect is whether or not the changes observed in the transcriptome are as a direct, or indirect consequence of the mutation. Caution must be applied if data are generated from 'non specific' structures – such as the whole embryo compared to tissue specific cells, as multiple genes can be expressed at the same time in different tissues. This could result in masking important information regarding tissue specific function. To highlight this point, in my case, Sox32 is expressed only in endodermal cells while Mixl1 is expressed both in mesodermal and endodermal cells.

In order to update the preliminary GRN model with data obtained from perturbation experiments, I focused first on the subcircuit network involving interactions of 4 key endodermal TFs (*sox32*, *sox17*, *gata5* and *mixl1*) (Figure 6.8). I added new interactions to this particular kernel of genes that I derived from overlapping information obtained from my genome-wide transcriptome perturbation experiments and the existing WT time series. In

mixl1 mutants, only *sox32* and *gata5* were statistically significant downregulated. This suggests that in WT embryos, a functional Mixl1 protein activates and promotes both *sox32* and *gata5* expression, while absence of the protein in the mutant embryos leads to reduced gene expression. In *sox32* mutants (Figure 6.8B), both *sox17* and *gata5* expression were downregulated whereas *mixl1* expression was upregulated, suggesting that Sox32 protein affects the expression of the other 3 genes in this subcircuit. This suggested a positive feedback of Sox32 on *sox17* and *gata5* in the WT, where Sox32 inhibited *mixl1* expression in the WT. Sox32 regulates *sox17* expression, and there are fewer endodermal cells present (as quantified by *sox17* expression) in the *mixl1* mutant, suggesting that *mixl1* indirectly, most likely through regulation of *sox32*, influences *sox17* expression, and therefore development of endodermal cells. It can be speculated that the reason *sox17* expression levels are not affected in the *mixl1* mutant is because of the time point at which the data for the RNA-seq was gathered (5.25 hpf); the reduced numbers of *sox17* positive cells only showed at a later developmental stage (9.00 hpf *in situ* staining). Additionally, the RNA-seq was performed in whole embryos, and as mentioned before, this can mask smaller tissue specific effects that fall below the threshold of significance due to high background noise. I would speculate that, if only endodermal cells were considered, RNA-seq of *mixl1* mutant cells would show a higher downregulation of *sox17*, even at the earlier time point of 5.25 hpf. Overall however, taken together, the results of the RNA-seq concur with the previously published literature (Kikuchi et al., 2000; Kikuchi et al., 2001; Aoki et al., 2002a; Chan et al., 2009a) and highlight the regulatory logic of *mixl1* feeding into *sox32,* which then regulates *sox17* (Figure 6.8 A).

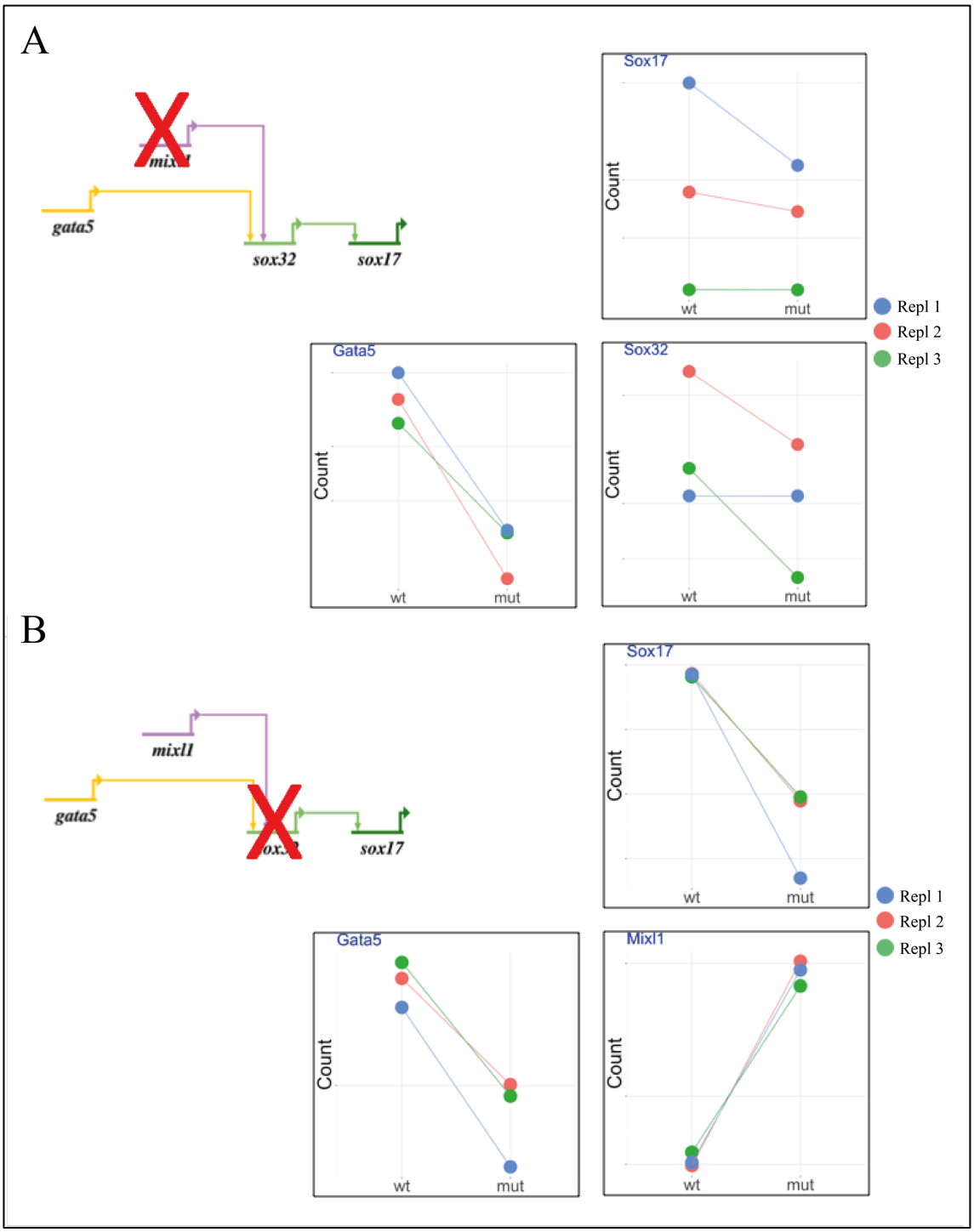


Figure 6.8 Construction of *sox32*, *sox17*, *gata5* and *mixl1* gene regulatory network. The interactions were determined by examining perturbation analyses in which one gene was perturbed and asking how its loss affected expression of other candidates. (**A**) Normalised read counts in *mixl1* mutant for *sox17, gata5* and *sox32* as labelled. The expression of both *gata5* and *sox32* in the absence of functional Mixl1 protein was statistically significant decreased suggesting a positive regulation of these two genes by Mixl1. No statistically significant change in the counts for *sox17* between the WT and the mutant was detected by DESeq2, which was further reinforced by additional RT-qPCR data. (**B**) Normalised read counts in the *sox32* mutant for *sox17, gata5* and *mixl1*. Non functional Sox32 protein affected the expression of all three genes; *sox17* and *gata5* were downregulated in the mutant suggesting positive regulation of Sox32 on these two genes in the WT. The

opposite relationship was seen for *mixl1* expression which was significantly upregulated in the mutant suggesting a role for Sox32 protein in repressing the expression of *mixl1* in the WT.

Next, I proceeded to integrate the information obtained from the mutants together with published time series gene dynamics, and a pattern was easily discernible. *mixl1* is expressed early in development, starting from the blastula stage, and is followed shortly thereafter by increased *sox32* expression. As *sox32* expression peaks during gastrulation, *mixl1* expression begins to decrease. This suggests that Mixl1 plays a role in turning on *sox32* expression. Sox32 then inhibits expression of *mixl1* in a classic negative feedback loop. These data also showed that *sox32* window of expression overlaps *gata5* expression during the gastrulation process, it is reasonable to postulate that Sox32 increases the expression of *gata5*, and that *gata5* subsequently drives expression of *sox32* - in a positive feedback loop with two mutual activators (Figure 6.9).

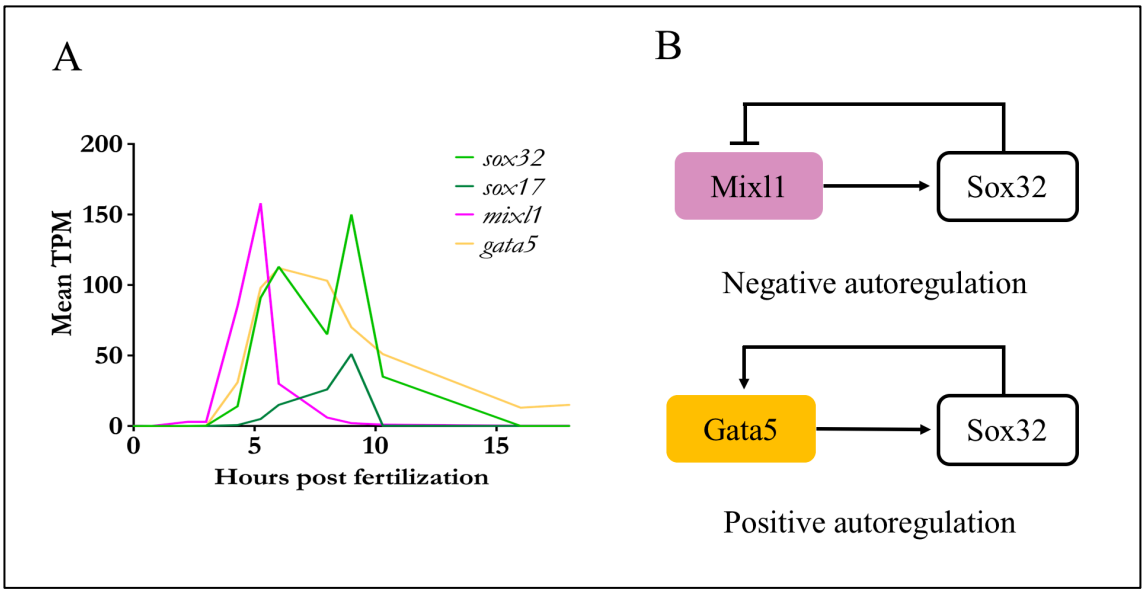


Figure 6.9 Positive and negative feedback loops in the *sox32*, *sox17*, *gata5* and *mixl1* kernel. (**A**) Dynamics of gene expression for *sox32*, *sox17*, *gata5* and *mixl1* transcripts during the first 15 hpf of development. By integrating the information from the transcriptome time series study with my RNA-seq analysis of mutant lines, I speculated (**B**) a negative autoregulation loop between *mixl1* and *sox32* and a positive autoregulation loop between *gata5* and *sox32*.

Altogether, combining these types of information allowed me to start adding new interactions around this kernel of endodermal TFs and the results are exemplified in Figure 6.10.

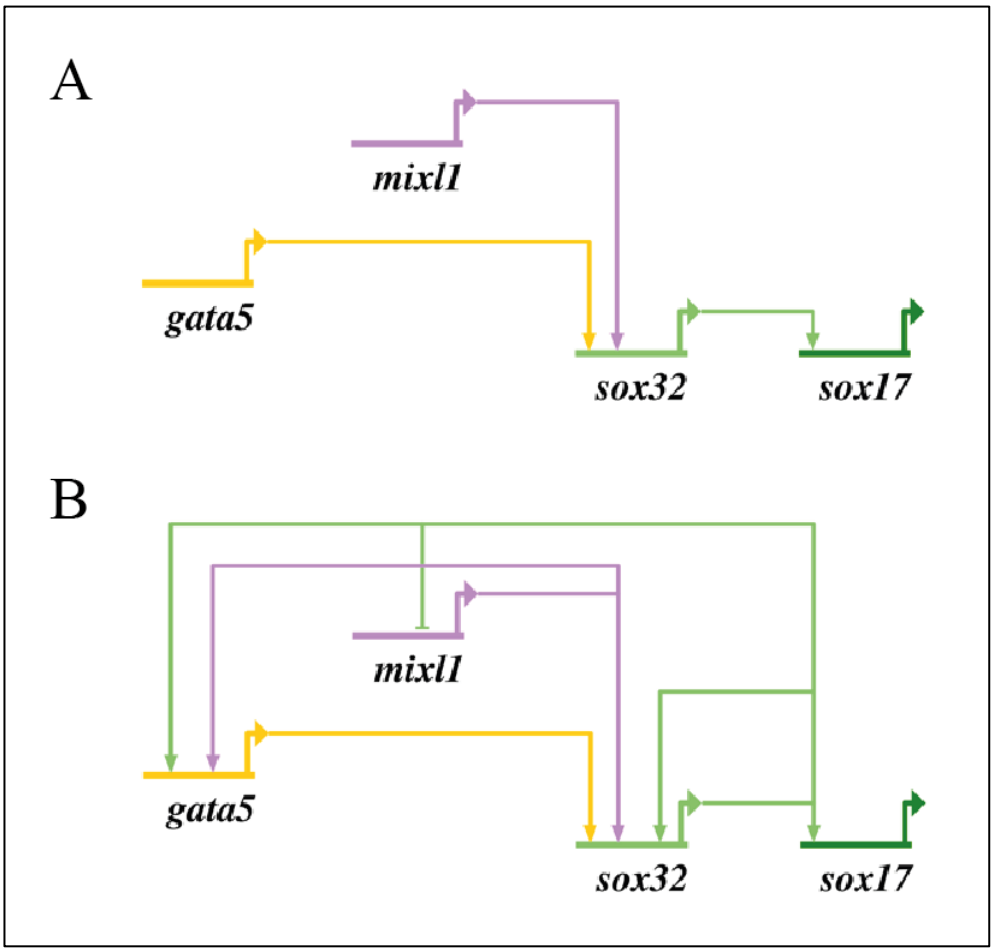


Figure 6.10 Gene regulatory networks based on Sox32 and Mixl1 perturbation. (**A**) Original and (**B**) updated subcircuits of the interactions between *sox32*, *sox17*, *gata5* and *mixl1* determined by combining WT time series datasets with mutant RNA-seq data as described in the text. Both *gata5* and *mixl1* are involved in the generation of endodermal cells at late blastula stages and also in the maintenance of endodermal *sox17* expression during gastrulation through the expression *sox32*.

Positive feedback loops in GRNs have been connected to the property of resilience of the biological system, as they increase the robustness and stability of the initial signal (Sharifi-Zarchi et al., 2015; Roy et al., 2017). The interaction between *sox32* and *gata5*, supports the ability of the system to be activated in response to a small input signal, as the input signal is then both sustained and amplified significantly due to the feedback loop (Mangan and Alon, 2003; Mitrophanov and Groisman, 2008; Ahnert and Fink, 2016). This is in accordance with Kikuchi et al., (2001) where endodermal precursors showed a cell autonomous commitment towards endodermal fate simply by expressing only *sox32*. Low levels of Sox32 might not be enough for cells to commit to an endodermal fate, but the positive feedback loop that I am proposing would mean that *sox32* then turns on *gata5* expression (and realistically other TFs), which in turn then increases *sox32* expression, until a level is reached that is high enough to

induce *sox17* expression – and therefore commitment to the endodermal fate. This claim is further substantiated by the work of Reiter et al., who, in 2001 showed that endodermal cells in *gata5* mutants showed lower levels of *sox17* expression, and that the maintenance of *sox17* expression via *gata5* is mediated by *sox32*.

Parallel to the positive feedback loops around *sox32* that reinforce the endodermal fate of a cell, the interaction between *mixl1* and *sox32* is also important for definitive commitment to endodermal fate. The literature suggests that *sox32* expression is driven by *mixl1* and other mesendodermal TFs at the midblastula stage, and my RNA-seq and ChIP-exo data suggests that at later stages during endodermal specification Sox32 physically interacts with the *mixl1* promoter - preventing *mixl1* transcription. This outcome advocates that Sox32 represses the expression of multiple early genes later in development. Data derived from murine embryonic stem cells shows that constitutively active expression of *Mixl1* causes increased commitment of cells to the endodermal lineage (Lim et al., 2009) – if a similar mechanism holds true for zebrafish, then switching off *mixl1* transcription by Sox32 at the onset of endodermal commitment could be an important mechanism in maintaining the correct balance between numbers of mesodermal and endodermal cells required for successful development.

To gain further insight into the feedback loop between Mixl1 and Sox32, I performed a *sox32* RT-qPCR in *mixl1* mutants and a *mixl1* RT-qPCR in *sox32* mutants. I then compared the respective expression levels to that of WT embryos (Figure 6.11). These data verified that *sox32* expression was transiently driven by Mixl1 at early stages of development and in addition, ChIP-qPCR also confirmed that Sox32 binds upstream of the *mixl1* promoter (see Figure 6.15), providing further evidence that Sox32 is directly controlling *mixl1* expression, as already confirmed at earlier stages by Nelson et al. (2017).

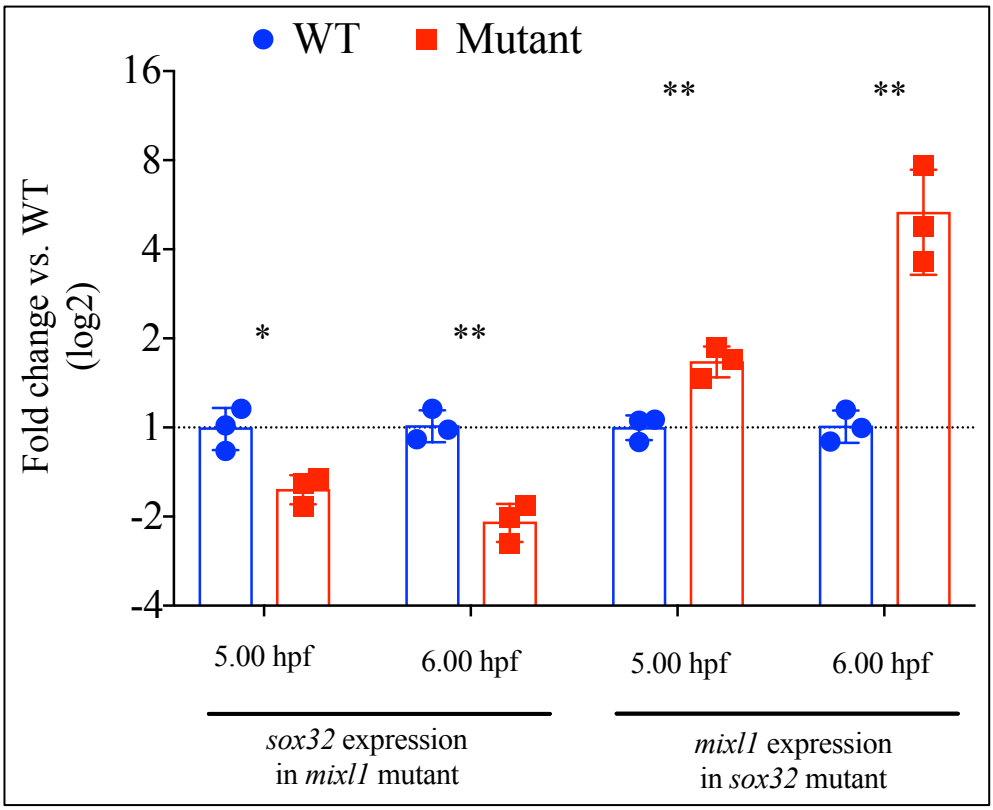


Figure 6.11 Expression of *sox32* and *mixl1* in the *mixl1* and *sox32* mutants respectively. *sox32* was downregulated in the *mixl1* mutant while *mixl1* expression was increased in the *sox32* mutant. We can deduce therefore that in WT embryos, Mixl1 positively regulates *sox32* expression and Sox32 reduces *mixl1* expression. As part of this feedback loop, Mixl1 both drove and sustained *sox32* expression, as *sox32* expression was reduced in *mixl1* mutants. Data are represented as mean ± SEM (n = 3) and fold change is displayed relative to WT expression. Unpaired two-tailed t test *$p \leq 0.05$, **$p \leq 0.01$.

It is worth noting that *sox32* is only expressed in endodermal cells, however *mixl1* is also expressed in other cell types and therefore gene expression quantification of *mixl1* using whole embryos does not depict the dynamics in endodermal cells. As informative as these RT-qPCRs were, they do not provide definitive proof of the suggested feedback loop and further experiments to quantify transcript levels in the mutant are needed. Spatiotemporal mapping of expression patterns together with perturbation assays provides extensive information, however direct evidence of physical interactions between identified TFs and the regulatory regions of downstream gene candidates is necessary to build a GRN.

There are several different techniques available to detect protein–DNA interactions including chromatin immunoprecipitation (ChIP), gel electromobility shift assay (EMSA) and DNase footprinting (Elnitski et al., 2006; Dey et al., 2012). Reporter gene assays, where the putative regulatory sequence is mutated can also be used to assess the direct interaction with

a TF. In particular, ChIP-seq, where ChIP is coupled with high-throughput sequencing, has substantially increased the ability to identify direct target genes *in vivo* (Johnson et al., 2007; Schmidt et al., 2009). Although large datasets of physical connections have been generated, the vast majority of these regulatory connections have not been validated, as this requires laborious and time-consuming mutagenesis assays to verify the suggested physical interaction. Another validation approach is to confirm putative connections through computational searching of the conserved non-coding region followed by functional analysis for further validation (Chan et al., 2009a; Nelson and Wardle, 2013; Bhatia et al., 2014; Nash et al., 2017).

The main drawback with ChIP techniques is the requirement for antibodies specific to the protein of interest, which are difficult to obtain for some species such as zebrafish (Nelson et al., 2014; Wardle and Tan, 2015) or are not sufficiently specific, as described in Chapter 3. Nevertheless, several studies have highlighted the power of ChIP-seq assays in different model organisms. In *Xenopus,* genome-wide binding of several TFs interplaying in both endoderm and mesoderm formation have been studied: Otx2, T-box TFs, Smad2/3, Foxh1, Lim1 and Gsc (Gentsch et al., 2013; Yasuoka et al., 2014; Gentsch et al., 2015; Charney et al., 2017a). In zebrafish, Morley et al., (2009) contributed by describing the GRN directed by Tbxta during mesoderm formation and patterning in the early zebrafish embryo by ChIP-ChIP. They were able to identify direct downstream gene targets (*noto*, *tbx6*, *tbx16*) and highlighted the diversified role of Tbxta in an array of developmental functions such as notochord specification, muscle specification and L-R patterning. ChIP-seq was also successfully used to identify binding of Pou5f3, a homolog of the mammalian pluripotency TF Oct4, and binding of SoxB1, which both overlap with genes first zygotically expressed in the zebrafish embryo (Leichsenring et al., 2013). Additionally, ChIP-seq analysis of Nanog, Mxtx2, Eomesa, Smad2 and Ldb2a revealed dynamic changes associated with promoter and enhancer activities and exposed a conserved transcriptional network in early mesendoderm induction in the early blastula (Liu et al., 2011; Xu et al., 2012; Nelson et al., 2014). The Nelson et al. paper from 2017 partially incorporated these data to enhance the understanding of the mesendodermal network, but even then, not all publicly available data had been put collated and integrated into a comprehensive GRN, which is what I attempted to do in my PhD.

I next focused on the particular subcircuit described previously; information was still missing regarding whether the interactions between Mixl1 and its targets *sox32* and *gata5*

were direct, or indirect, for example via another TF and if Sox32 directly bind *mixl1* and *gata5*. (Figure 6.12).

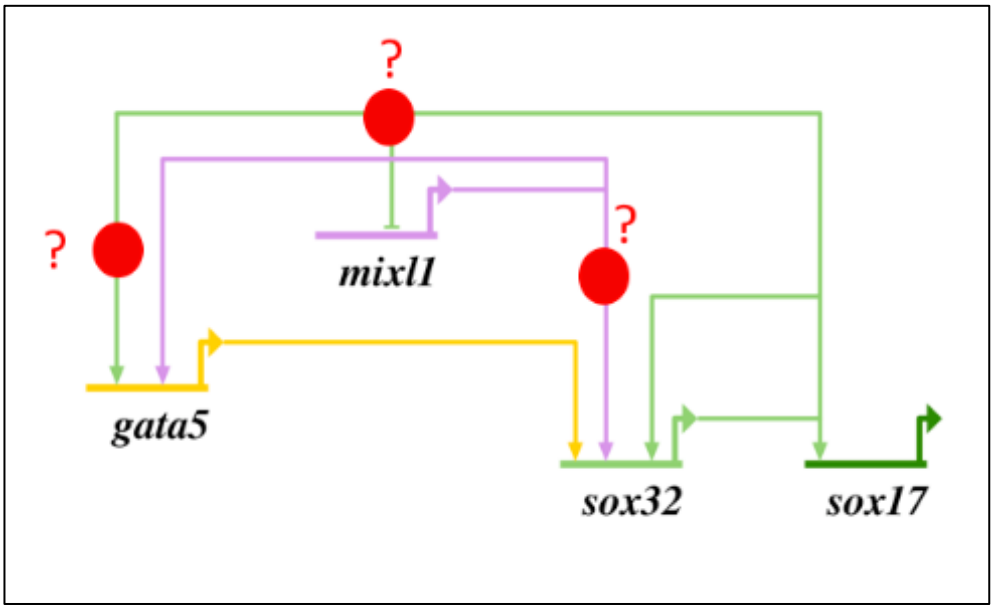


Figure 6.12 GRN subcircuit of the interactions between the 4 TFs Gata5, Mixl1, Sox32 and Sox17. Red circles represent regulatory events where interactions could be direct or indirect. By using RNA-seq in the mutant zebrafish, I identified a set of downregulated genes giving some insight into these interactions, however further validations were necessary to define whether the interactions between these TFs were direct or indirect.

To validate the putative targets of the subcircuit, I investigated both the published Mixl1 ChIP-seq dataset (Nelson et al., 2017) and my Mixl1 and Sox32 ChIP-exo datasets. From the analysis, I identified that both Sox32 and Mixl1 did indeed directly bind upstream of both *mixl1* (Figure 6.13*)* and *gata5* (Figure 6.14), confirming that these regulatory interactions were direct.

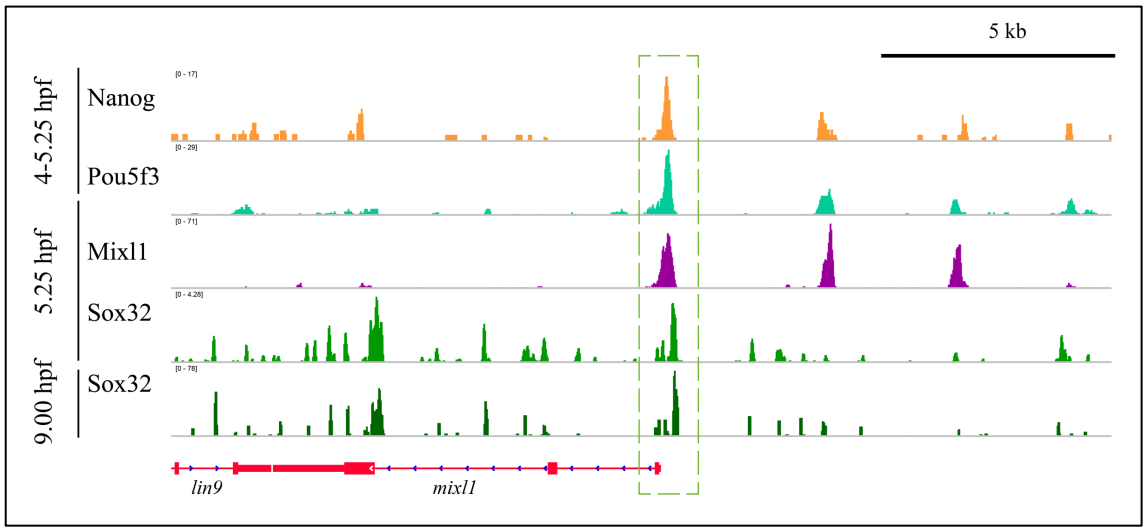


Figure 6.13 Sox32, Mixl1, Pou5f3, Nanog and Mxtx2 ChIP-exo/ChIP-seq at indicated developmental stages proximal to *mixl1*. Peak heights in reads per million (RPM) are reported. Pou5f3, Nanog and Sox32 form complexes to drive mesendoderm patterning. Boxed region indicates peak used for ChIP-qPCR validation.

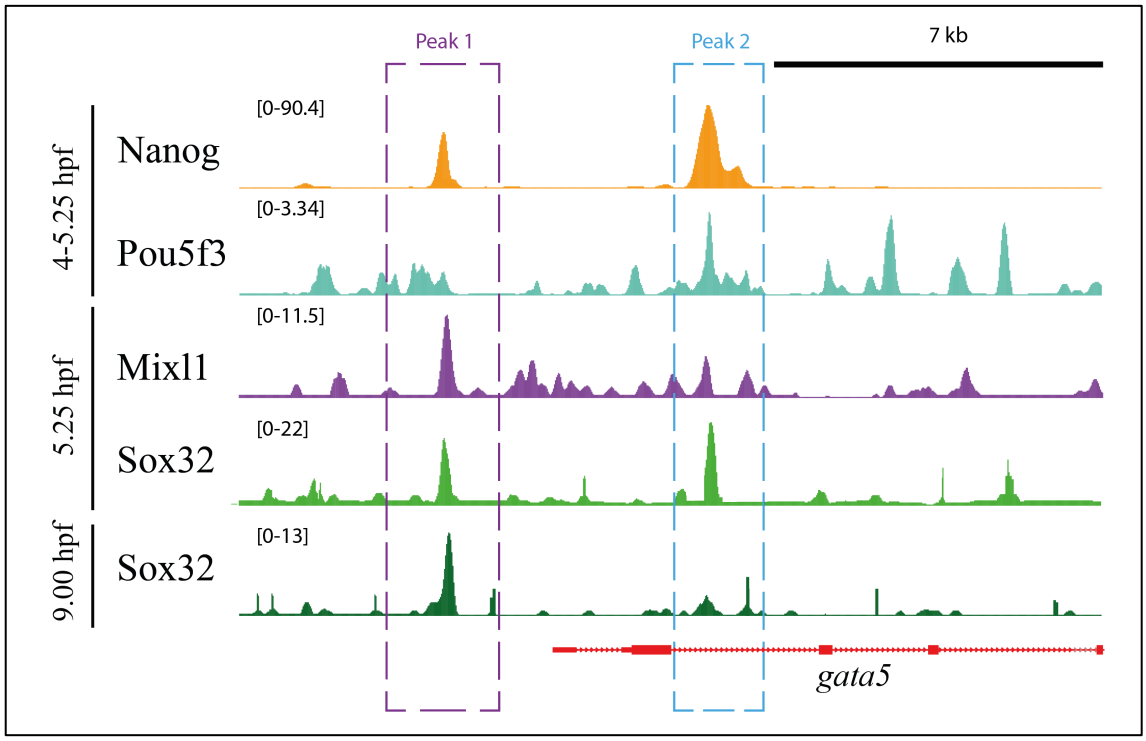


Figure 6.14 Stage-matched Sox32, Mixl1, Pou5f3, Nanog, and Mxtx2 ChIP-exo/seq at *gata5* genomic locus. Peak heights in reads per million (RPM) are indicated. Pou5f3, Nanog and Sox32 form complexes to drive mesendoderm patterning. Two boxed regions indicate peaks used for ChIP-qPCR validation.

ChIP-qPCR validation confirmed time specific Sox32 and Mixl1 binding at both *mixl1* and *gata5* (Figure 6.15). Thus, consistent with molecular interaction previously shown by RNA-seq analysis (Figure 6.10B), ChIP technique proved direct binding events of Sox32 and Mixl1 in the upstream regions of *mixl1* and *gata5*.

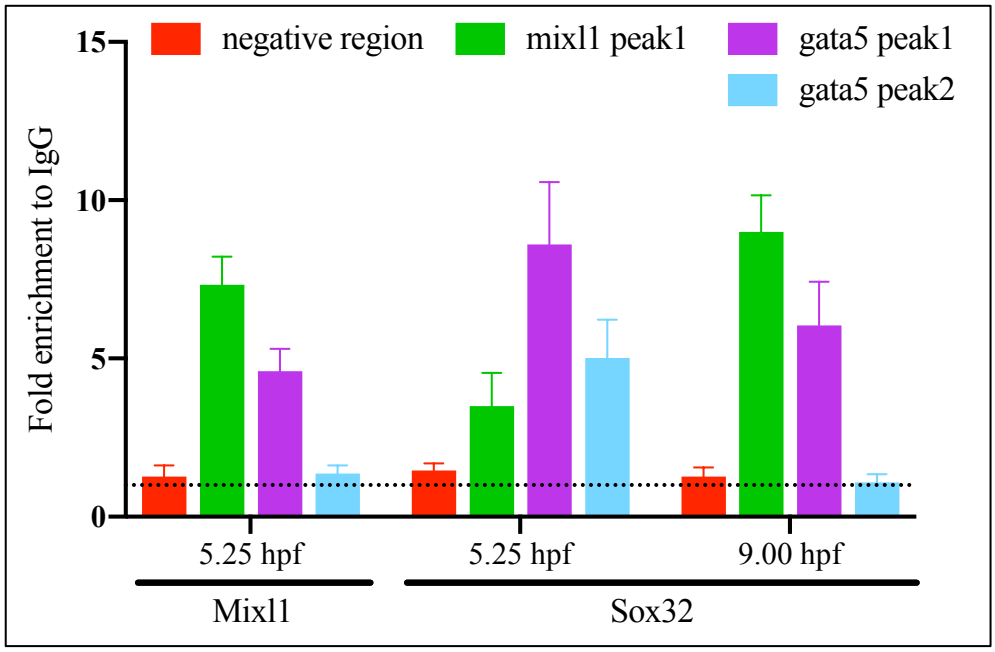


Figure 6.15 ChIP-qPCR validation of Sox32 and Mixl1. Clear enrichment over IgG-negative control (dotted line) was observed for both Mixl1 and Sox32 ChIP samples analysed. Specifically, high levels of enrichment for Mixl1 in the upstream region of Mixl1 and Gata5 (peak 1) at the 5.25 hpf and enrichment for Sox32 at the transcriptional start of Mixl1 and Gata5 (peak 1) at both 5.25 and 9.00 hpf were observed. Gata5 peak 2 showed no enrichment for Mixl1 and Sox32 at 5.25 hpf. Data are shown for 2 developmental stages, normalised over negative control region and input and colour coded per boxed regions as shown. Error bars indicate SEM from 2 replicates.

I then extended the analysis to other genes that I identified with my RNA-seq analysis, including Sox32 downstream genes: *txn*, *prdx5*, *cdx4* and *nanog* (Figure 6.16).

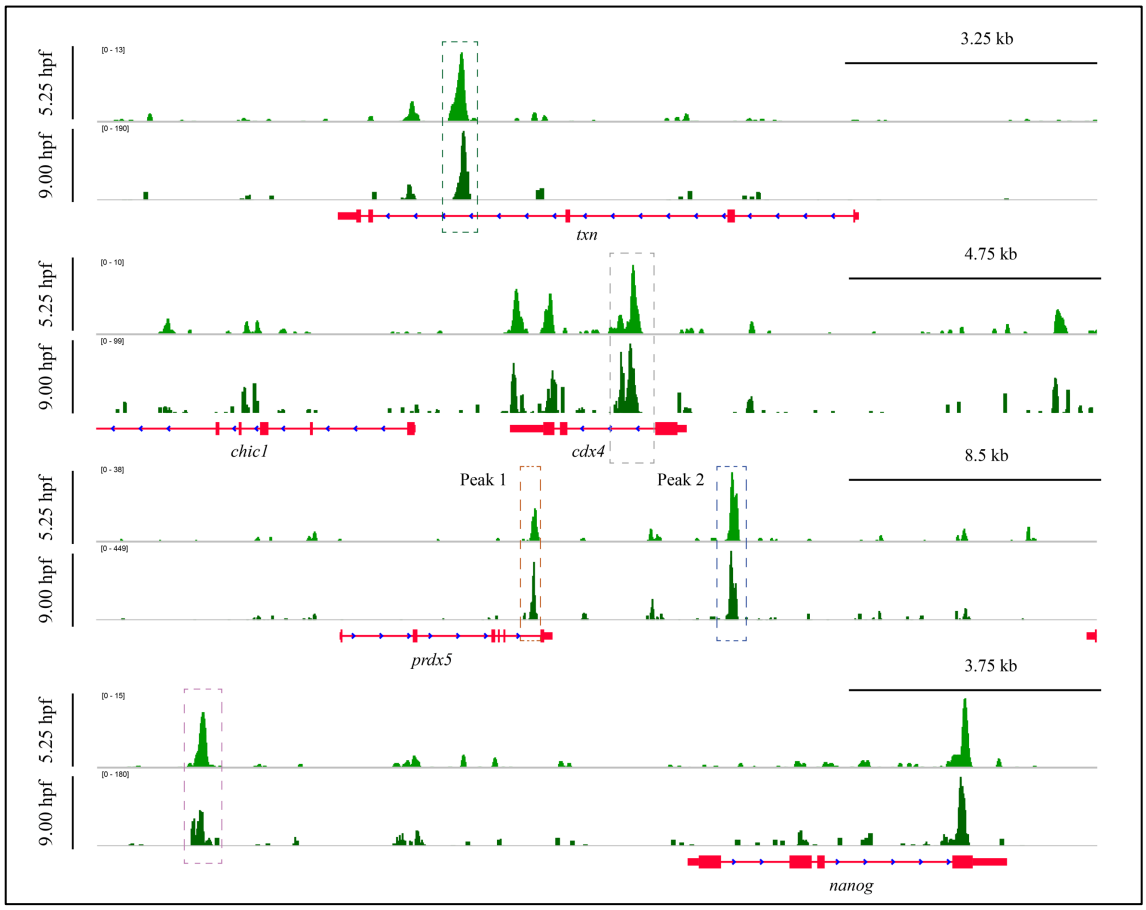


Figure 6.16 Sox32 binding to *sox32* mutant DEGs. Genome browser view of peaks on multiple genes identified by RNA-seq analysis on *sox32* mutant. Pharyngeal arch marker *txn*, pronephros development gene *prdx5*, pancreatic marker *cdx4* and Nodal signalling activator *nanog*. Peak heights in reads per million (RPM) are indicated. Boxes indicate statistically significant Sox32 peaks used for ChIP-qPCR validation.

I was also interested to confirm whether Sox32 and Mixl1 were binding the endodermal genes *dusp4* and the mesodermal genes *tbxta* and *dlc*. As shown in Figure 6.17, putative binding site were identified upstream of all 3 genes.

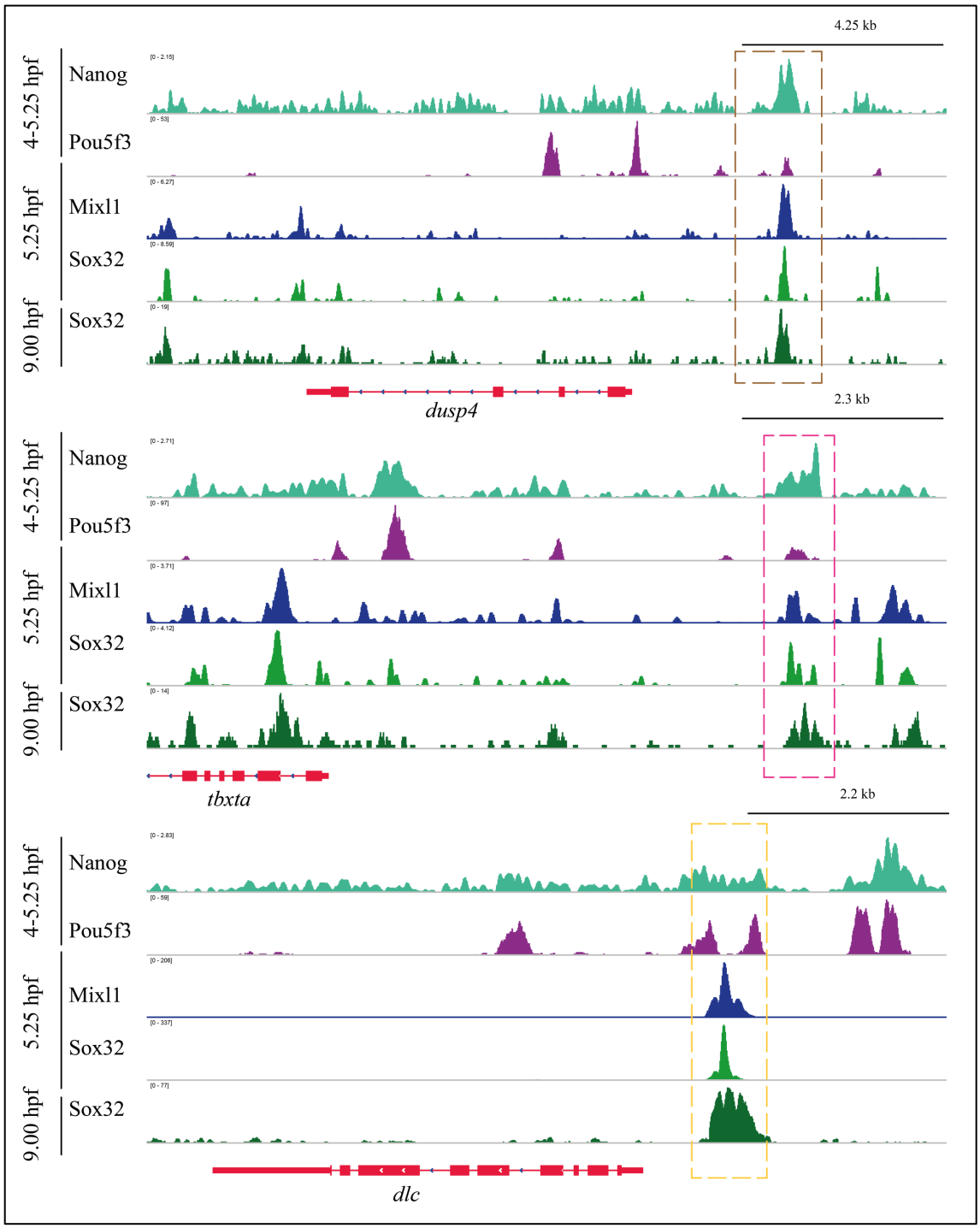


Figure 6.17 Sox32 and Mixl1 chromatin binding to endoderm and mesoderm regulated genes. Genome browser view of peaks on Nodal regulator gene *dusp4* and mesodermal genes *tbxta* and *dlc*. Peak heights in reads per million (RPM) are indicated. Boxes indicate peaks used for ChIP-qPCR validation.

To validate the ChIP-exo results I then performed ChIP-qPCR on independent biological samples for all the selected regions (see Chapter 3 Methods). As shown in Figure 6.18 the peaks I tested had robust enrichment (from 2.4 to 8.9-fold) over the IgG ChIP control.

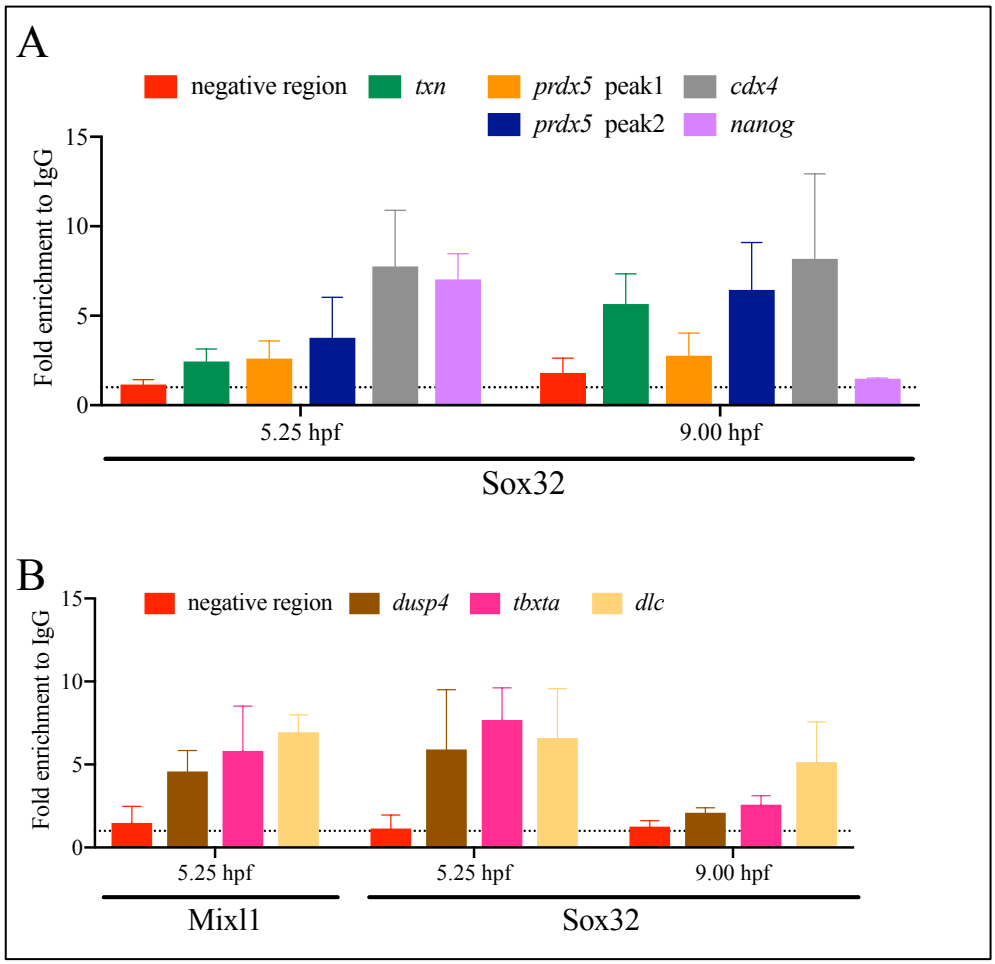


Figure 6.18 ChIP-qPCR validation of Sox32 and Mixl1 target genes at the indicated stages. (**A**) Sox32 ChIP-qPCR showed binding to promoter region of *txn*, *prdx5*, *cdx4* and *nanog* at 2 developmental stages. (**B**) ChIP-qPCR showed Sox32 and Mixl1 binding to promoter region of endodermal gene *dusp4* and mesodermal genes *tbxta* and *dlc* at 5.25 hpf. *dlc* was also bound at 9.00h hpf by Sox32. Data are represented as mean ± SEM relative to IgG control (dotted line) (n = 2). Colour coded per peak as in earlier figures.

These experiments not only showed that the ChIP-exo data were concordant with independent ChIP-qPCR data but also proved that my ChIP-exo dataset can be use to test specific binding events upstream of endodermal and mesodermal genes. From this analysis I derived that seven more genes were likely to be part of the GRN as were directly regulated by either Sox32 and/or Mixl1.

6.3 The updated endoderm GRN

Cis-regulatory elements (enhancers or silencers) coordinate and control the expression of genes encoding TFs and cell signaling pathway components, generating precise genetic cascades during development, reviewed in Davidson et al. (2002), Levine and Davidson (2005) and Peter and Davidson (2011a). GRNs present a different approach to illustrate these

developmental pathways as a genomic network topology, rather than a cascade of inter/intracellular processes. So far in this chapter, I have reviewed the available biological data that I needed to generate a more comprehensive GRN; now I merge this information together to construct the GRN.

Endoderm formation in zebrafish is patterned by the combinatorial effect of 3 major signalling pathways Nodal, Fgf and Bmp. These signals generate temporal and positional cues that shape the future of the cells by constricting the expression of TFs to within specialised cell populations. Nodal acts positively and both Bmp and Fgf act negatively on the formation of endoderm precursors. Bmp signalling restricts endodermal precursor cells on the ventral side while Fgf in the dorsal margin coregulates Nodal target genes, restricting the number of endodermal progenitors (Mizoguchi et al., 2006; Poulain et al., 2006; van Boxtel et al., 2018). The YSL is an important structure in early zebrafish embryos as it acts as a signalling centre that controls morphogenesis and regulates mesoderm/endoderm pattering. Maternal TFs and signalling ligands activate Nodal signalling which then regulate *ndr1* and *ndr2* (Bjornson et al., 2005; Bruce et al., 2005; Xu et al., 2012; Xu et al., 2014). These are inducers of mesoderm and endodermal commitment which in turn activate expression of *sox32* and other downstream endoderm TFs (Reiter et al., 1999; Rodaway et al., 1999; Kikuchi et al., 2000; Poulain and Lepage, 2002; Poulain et al., 2006; Chan et al., 2009a; Tseng et al., 2011; Liu et al., 2018). Nodal acts through Acvr1ba receptors, Tgdf1 coreceptors and intracellular signal transducers Smad2/Smad4, and the consequence of these signalling cascades is context dependent (David and Rosa, 2001; Aoki et al., 2002b; Liu et al., 2018).

The combination of all the information described in this chapter has allowed me to increase the number of connections in the previously published GRN, a static view of this BioTapestry network is shown in Figure 6.19. This systems level perspective provides a summary of all the inputs affecting each gene with all connections visualised at once, regardless of time and space. This network includes 43 TFs and 14 growth factors, with genes distributed vertically based on approximate activation time, thus representing the chronology of expression. It is an expansion of the previous network of Chan et al. (2009) and Nelson et al. (2017). This network covers a timespan of approximately 5 hrs, at the top a wide variety of zebrafish maternal and zygotic regulatory factors and signalling ligands have been described. Together, these ligands and factors initiate mesendoderm formation in the midblastula, with the beginning of Nodal signalling in the YSL. During this time of rapid developmental, morphogenetic movements

give rise to the different germ layers, and the interplay between TF binding and signalling molecules activity contributes to the acquisition of endodermal cell fate (Figure 6.19).

Early mesendodermal specification is the result of 3 coordinated signals: Nodal (from vegetal pole to animal), Fgf signals on the dorsal side, and Bmp signals on the ventral side. Maternal Nodal ligand Ndr1 and maternal TFs activate genes expression by first acting on the YSL which induces cell autonomous Nodal signalling through the production of zygotic *ndr1* and *ndr2*. *mxtx2* and *nanog* also contribute to the activation of this loop and the regulation of *mir-430.* Nodal induces long-range Fgf signalling whilst concurrently inducing the Fgf signalling inhibitor Dusp4, which inhibits Fgf signalling closest to the YSL. Nodal ligands secreted by the YSL also turn on expression of *ndr1/2* and *left1/2* which are also then regulated by the action of *mir-430* and work to enhance the action of *dusp4* in specifying endodermal fate in cells closer to the margin. Maternal and zygotic TFs such as Foxh1, Pou5f1 and Eomes play important roles together with Nodal signalling, to activate *smad2/3* and allow the transcriptional activation of the earliest zygotic mesendodermal TFs including *sebox*, *mixl1*, *gata5*, and *sox32*. Nodal signalling, through long range Fgf signalling, also activates mesoderm specific TFs such *tbxta*, *noto*, *tbx16*, *lhx1a*, *bhik* and *gsc*. This is also helped by *lft1* and *lft2* which were blocked by *mir-430* when initiated earlier by Nodal. By midblastula stage, mir-430 is degraded, Lft1 and Lft2 are released and are now able to inhibit extracellular diffusion of Nodal signalling at the receptor level in cells further away from the margin and reinforce Fgf signalling. At the onset of gastrulation (5.25 hpf) the coexistence of both endoderm and mesoderm TFs activated by Nodal indicates that the cells have the potential to follow either lineage. *sebox*, *mixl1*, *eomesa*, *gata5* and *gata6* create positive feedback on *sox32* and *sox17* which also autoregulate themselves, creating a stable lock-on system to safeguard the endoderm uniformity of expression within the cells. Genes such *foxa2* also feedback in this motif as they are activated by s*ox32*, *gata6* and *otx2. otx2* also activates *sox17* expression, and *sox32* activates itself, shut down *nanog* and *mxtx2* expression at this stage. During gastrulation the *sox32*/*sox17* lock turns on/off the expression of mesodermal TFs and positively regulates genes associated with cell migration and later stages of pharyngeal, pancreatic and liver specification.

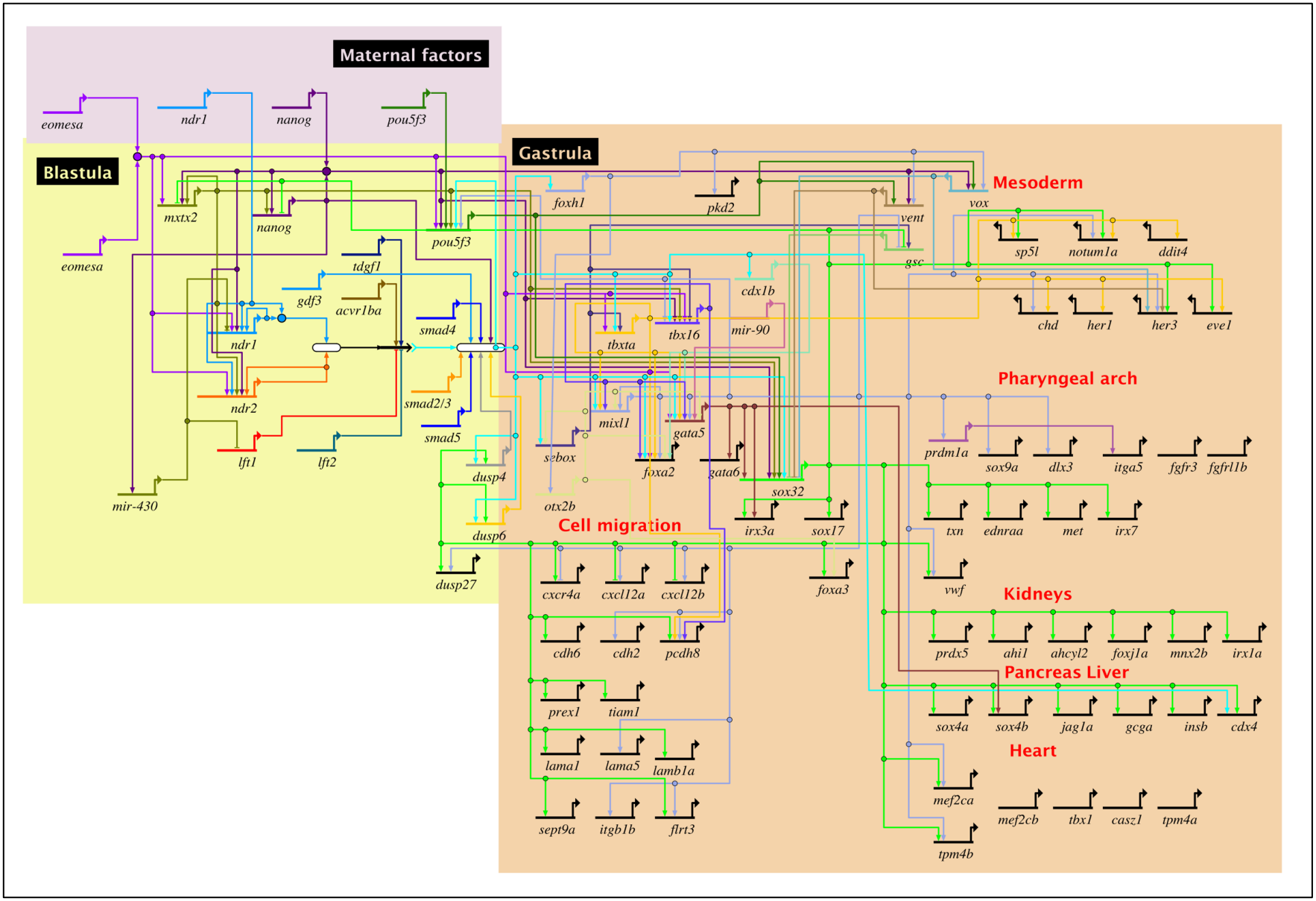

Maternal factors
Blastula
Gastrula
Mesoderm
Pharyngeal arch
Cell migration
Kidneys
Pancreas Liver
Heart
eomesa
ndr1
nanog
pou5f3
mxtx2
foxh1
pkd2
vox
vent
gsc
sp5l
notum1a
ddit4
chd
her1
her3
eve1
tdgf1
gdf3
acvr1ba
smad4
smad2/3
smad5
ndr2
lft1
lft2
mir-430
cdx1b
mir-90
tbx16
tbxta
mixl1
gata5
foxa2
gata6
sebox
otx2b
sox32
irx3a
sox17
foxa3
dusp4
dusp6
dusp27
prdm1a
sox9a
dlx3
itga5
fgfr3
fgfrl1b
txn
ednraa
met
irx7
vwf
cxcr4a
cxcl12a
cxcl12b
cdh6
cdh2
pcdh8
prex1
tiam1
lama1
lama5
lamb1a
sept9a
itgb1b
flrt3
prdx5
ahi1
ahcyl2
foxj1a
mnx2b
irx1a
sox4a
sox4b
jag1a
gcga
insb
cdx4
mef2ca
mef2cb
tbx1
casz1
tpm4a
tpm4b

Figure 6.19 Zebrafish mesendoderm GRN from early blastula through to late gastrula that highlights the complexity of the transcriptional networks operating during endoderm formation. At the top of the network sits maternal TFs and signalling ligands. All targets of signalling pathways are connected through » indicating cell surface receptors to their respective intracellular signal transducers/TFs. Temporal information is provided from top to bottom. Maternal signalling ligand (Ndr1) and TFs (Eomes, Pou5f3) are connected back to the same zygotic signal through •, so that all connections from a given pathway feed through a single node.

6.3.1 Compartmentalisation of endoderm lineages in different motifs

Gene regulation networks are composed of small sets of recurring interaction patterns called network motifs which are elementary circuits for signal propagation and guarantee robust signal propagation of the pathways (Milo et al., 2002; Chan et al., 2009b; Roy et al., 2017). Each motif supports specific dynamic functions not only by forming dense clusters of *cis*-regulatory elements, but also by creating independent paths which serve to preserve their autonomous functions even in the presence of perturbation. Deconstructing the network structures into smaller units also reduces the inherent complexity and helps in inferring new interactions. It allows us to observe how GRNs contain repeated occurrences of the same loop. It reflects the structural nature of biological processes, as these consist of pathways that mainly act both on their own and in crosstalk with each other in a context dependent manner to generate a coordinated output.

The increased number of connections between genes allows the identification of more regulatory motifs in the updated GRN and particularly highlights the important roles that autoregulatory loops, feedback loops and feedforward loops play in influencing the architecture of the network. Autoregulatory loops encompass self-regulation of a TF or signalling pathway; positive feedback loops enable continuous expression of genes that are critical for the same lineage commitment. Negative feedback loops are a self-regulating system in which increased output from the system inhibits its own activity in order to maintain an ideal steady state.

Current evidence supports three functionally relevant modules according to territory and time, that are important building blocks in zebrafish endodermal GRNs. The first one is establishing appropriate levels of Nodal signalling to induce the mesendoderm cell population which involves both positive and negative feedback loops, coupled and driven by the same signal transduction components. The second module describes the activity of an induced Nodal

related set of TFs that establish endodermal and mesodermal boundaries and ensure the activation of sufficient levels of *sox32*/*sox17*. The third and final module describes the gene regulatory events around *sox32*/*sox17* that lead to endoderm specification, migration and patterning of endoderm related tissues.

6.3.2 Functional motifs uncover the robustness of the GRNs

Using the pipeline described above (literature-based, data mining, temporal and spatial expression domains, perturbation assays, RNA-seq and ChIP-seq), I present a model of the GRN underpinning zebrafish mesendoderm development from the midblastula stage to the end of gastrulation (Figure 6.19). This updated network contains a total of 104 direct connections − 91 positives and 13 negatives. Combining the connections in my network with two previous GRNs (Chan et al., 2009b; Nelson et al., 2017) has resulted in a major improvement to the current network allowing the identification of both an internal lock-on system and additional *bona fide* direct connections between TFs.

From the analysis of these subcircuits and correlated internal loops, the most important properties that emerged from the GRN were the robustness of the system and the multiple lock-on systems at different stages to buffer functionality in the face of perturbation. This small set of recurring interaction patterns (network motifs) perform specific dynamic functions, with each subcircuit acting as a separate building block in creating, steadying and shaping the endoderm and preserving their autonomous functions.

Historically, perturbation experiments have been the key approach for dissecting and reconstructing a GRN; understanding the differences in GRNs under varying conditions (knockdown and overexpression) allows us to understand condition specific gene regulation, pinpointing the underlying logic behind the interactions (Li and Davidson, 2009). In accordance with the network, perturbation techniques that act at different levels of the pathway can output overlapping phenotypes due to compensation by other factors on the same level. For example, single *ndr1* or *ndr2* mutants form most of the endodermal structures whereas double mutants completely fail to form endoderm and dorsal mesoderm. In addition, the models show that *ndr1* mutants behave with a different time response to endoderm formation than *ndr2*. Single *ndr1/2* mutants still have maternal contribution from *ndr1*, and therefore the initiation of the endodermal pathway is only delayed, suggesting that they are capable of at least partially compensating the function of the other (Feldman et al., 2000; Dougan et al.,

2003). In a similar way, single perturbation of a mesendodermal TF (module 2 as described above) such as *gata5* or *mixl1* presents a weaker phenotype than the *sox32* mutant (Reiter et al., 1999; Reiter et al., 2001). In the latter, the *sox32*/*sox17* lock-in system cannot be activated and therefore non functional Sox32 protein does not allow expression of all the downstream genes shown in Figure 6.19. The absence of these genes (*casz1, mef2cb, tpm4b*, *foxa2, foxa3)* causes the fish to present with abnormal endodermal derivates such as gut tube malformations and cardia bifida. When only *gata5* or *mixl1* are not expressed, the *cis*-regulatory module can still activate output of *sox32*, which although reduced in activity still leads to the formation of endodermal structures. Each and every member of one motif shares a common TF regulating output, and this coregulation of the *cis*-regulatory elements confers robustness to the cluster.

It is the functional redundancy between these factors that explains the low, mild and severe phenotypes associates with different mutants. Gata5 mutants present a 10% reduction in endodermal cell numbers and cardia bifida (Reiter et al., 2001). *mixl1* mutants show a 60% reduction in endodermal cell numbers, a reduction in prechordal plate cell numbers and cardia bifida (Kikuchi et al., 2000). Only when the whole CRM is drastically modified will the GRN collapse, for example injecting *mixl1* mutants with *sebox* morpholino, the other member of the Mix homeobox gene family in zebrafish, increases the phenotype severity with abolishment of the prechordal plate and no endodermal cells (Poulain and Lepage, 2002). This cumulative effect is in line with the observations of Nelson et al. (2017), where *tbxta/tbx16* double mutants show a substantial loss of endoderm compared to single KO.

If the autoregulatory lock between *sox32* and *sox17* is not established, the whole endodermal fate is crushed. Once Nodal signalling activates *sox32,* it keeps activating itself and activates *sox17* from 5.25 hpf. This positive loop causes potential endodermal cells to rapidly accumulate the *sox32/sox17* transcripts and reach the 'threshold' of becoming endodermal. This loop is still not enough to ensure the correct cell migration necessary for endodermal patterning. Explant experiments have provided evidence in respect of this missing link. *sox32* mutant cells do not display any movement when transplanted into the animal pole, they still express the endodermal marker *sox17* but fail to translocate and therefore cannot reach the surface of the YSL and they are re-specified towards animal pole fates (Peyrieras et al., 1998; Dickmeis et al., 2001). When *acvr1ba* induced endodermal cells (a receptor of Nodal signalling) are transplanted into the same position, they can ingress into the inner layer of the embryo and by the end of segmentation at 24 hpf, they are integrated into endodermal

derivatives such as the pharynx, thus demonstrating that full activation of the Nodal pathway is sufficient to commit cells to an endodermal fate. Liu et al. (2018) further explored this observation, providing evidence that only cells expressing both *sox32* and the *acvr1ba* receptor can migrate and create endodermal structures. *acvr1ba* expressing cells injected with a *sox32* MO fail to internalize after being transplanted into the animal pole. Nodal signalling is therefore necessary and sufficient to initiate the internalisation process of endodermal cells, with its ligands acting in an autocrine fashion to initiate endodermal cell sorting. However, cells from a *sox32* mutant, in which the *sox32*/*sox17* motif is not active in the margin before gastrulation, fail to express downstream genes that regulate the development of endoderm derived organs, explaining the presence of cardia bifida. These results advocate that cells require the *sox32*/*sox17* motif to commit to endoderm, but also need additional signalling downstream of *acvr1ba* to correctly migrate. Only when these two requisites are met simultaneously cells are channelled to endodermal internalisation, cell migration and fate commitment. These observations also suggest the presence of as yet unknown molecular cues present in the network, which should be addressed in future work. As the authors point out, it will be interesting to examine further the interplay between Nodal, the receptor Acvr1ba, the coreceptor Tdgf1 and Smads, the main signal transducers. The system could be further expanded to include the role of Fgf signalling and the contribution of Dusp4. Moreover, my results highlighted the role of Sox32 in orchestrating a subset of migratory proteins, therefore integrating all these data will help us to better understand how endodermal cells process the overlapping signals of migration and commitment to become endoderm.

6.4 Summary of the chapter

Depicting early developmental processes as GRNs helps us to understand the complexity of the governing regulatory processes, especially as such processes hardly ever run in a linear fashion. Germ layer specification is determined by a myriad of TFs acting together in a complex fashion and it is hard to compare the level of evolutionary conservation between different species. Generating detailed GRNs is an important step towards a major goal in developmental biology; determining temporal and spatial expression patterns of all relevant genes during development. Combining the vast amounts of data generated from scRNA-seq datasets allows for temporal and spatial reconstruction of developmental processes; coupling this with recent advances in mathematical modelling means these data could potentially be used to create a comprehensive atlas of gene expression.

Identifying the core parts of GRNs regulating endodermal and mesodermal development is of special importance, as many human congenital diseases result from abnormal formation of these germ layers. Recent advances in stem cell research and regenerative medicine, combined with highly efficient *in vitro* differentiation techniques has allowed us to start understanding early developmental regulatory processes (Singh et al., 2015; Dogan, 2018). These advances have allowed us to generate a significant understanding of *in vivo* cellular differentiation programs, namely "GRN science". Comparing the GRNs generated in these human models with those derived from other model organisms such as *Xenopus* or zebrafish allows for the identification of evolutionarily conserved core networks, and subtle modifications of any/all members of these subcircuits might well allow researchers to identify disease-causing mutations (Emmert-Streib et al., 2014; Ober and Grapin-Botton, 2015; Charney et al., 2017b; Yiangou et al., 2018).

Previous work in zebrafish has generated vast amounts of data, and most of it has resulted from *in situ* hybridisation patterns observed in response to deleterious mutations or overexpression experiments. While these data indicate a certain dependency between factors, they do not prove any direct physical interaction between the identified players. This can lead to varying hypotheses ranging from a single, direct interaction to a complex network that may require many additional regulators. To compile our current knowledge of zebrafish development, I started building a GRN underlying early endoderm development by integrating dispersed data on expression, localisation and protein interactions. However, further data generation by promoter analysis, chromatin immunoprecipitation, as well as high throughput sequencing techniques were still needed to complete this picture.

Next generation sequencing has immense power in addressing questions regarding biological systems such as transcriptomes, chromatin accessibility and the study of protein-DNA interactions. The resolution of these techniques is additionally increased by the refinement of single cell sequencing methodologies that can identify interesting candidates that might previously have been masked by cell heterogeneity. The efficiency of these methods has improved massively over the last few years and combined with advanced mathematical and statistical approaches, has allowed the identification of likely regulatory relationships between genes based on their expression patterns, another step towards the refinement of existing GRNs.

The analysis and inference of GRNs requires specific algorithms that solve principally two challenging problems intrinsic to the nature of the data: the complexity of gene regulation mechanisms and the highly interconnected levels. Computational approaches integrating multisource biological data (RNA-seq, Chip-seq, ATAC-seq) in the last decade have started addressing this challenge (Gligorijević and Pržulj, 2015; Castro et al., 2019). Most approaches analyse the GRN at the transcriptional level and multiple models have been proposed for inference of GRNs from continuous or discrete approaches, static or dynamic interpolation, and quantitative or qualitative description. These include linear models, Boolean networks, Bayesian networks and recurrent neural networks (RNNs) (Liu, 2015). None of these approaches have yet been applied to study the endodermal GRN and future refinement of the endodermal GRN should focus on integrating the multi-complementary high-throughput datasets now available. Reconstruction of a top-down endoderm network by assembling all the data grouped by the DANIO-CODE consortium and inferring an endoderm transcriptional regulatory network seems feasible for future work. This type of analysis could help make specific predictions about the network properties and facilitate the design of downstream biological experiments; bioinformatics inferences are more time and cost-effective than wet lab research, and as they can make specific predictions about experimental outcomes, they should be used to inform experimental design.

Over the last two decades, our understanding of endoderm development has increased significantly, with various genetic screens in zebrafish identifying previously unknown genes. Similarly too, our understanding of the regulatory mechanisms that control correct endoderm formation has substantially improved. We have a good basic knowledge of the interactions that cause cells to adopt the endodermal fate at the beginning of gastrulation in the margin of the embryo. Nelson et al. (2017) used a combination of molecular biology and omics approaches to develop a GRN covering from late blastula to early gastrula stages, focusing on mesendodermal TFs. They showed new TFs interactions downstream of Nodal signalling and highlighted how specific genes can act as a switch in fate decisions, causing a cell to commit to one fate (endoderm) while simultaneously preventing its commitment to another (mesoderm). Understanding the importance of Nodal signalling is a great beginning, but the field does not yet completely understand the exact function(s) of other important signalling cascades. For example, Fgf signalling plays an important role in the induction of mesodermal genes over endodermal genes, but the exact mechanisms remain to be elucidated. Another example would be the importance of Dusp4 in endodermal specification of precursor cells;

Nodal signalling induces *dusp4* expression, but only cells positive for *dusp4* and *sox32* commit to the endodermal fate, while *dusp4*$^{+}$*sox32*$^{-}$ cells will not. By beginning of epiboly, most cells in the first 2 cell tiers express *dusp4*, but not all are specified as endoderm cells, suggesting that there are additional mechanisms in place to control endodermal commitment, but those are as yet not defined.

Refinement of any GRN is an iterative process, where the existing GRN is refined by multiple rounds of consolidation of existing data and integration of newly generated experimental evidence. In order to obtain useful data to refine this GRN, further experiments should put emphasis on collecting samples over a dense time series to zoom into this window of gastrulation development. This would allow us to identify additional candidate genes that might only be active over a short amount of time, and that have been overlooked by studies sampled over larger timeframes. These should then be confirmed by WISH and detailed establishment of spatial patterns of gene expression. A combination of promoter bashing and mutagenesis of the upstream regions of identified new candidate genes would then further refine our understanding of the gene dynamics during zebrafish endoderm development.

The overarching goal of my PhD was to apply these techniques to increase our understanding of the regulation of early endodermal development in zebrafish. Already existing data provided me with a foundation upon which I could build a more comprehensive GRN underpinning endoderm development; specifically, to establish a more complete subcircuit around Sox32 and Mixl1. I used chromatin immunoprecipitation coupled with deep sequencing (ChIP-exo) to identify genome wide Sox32 and Mixl1 binding sites. Defining these sites identified putative target genes and revealed possible regulatory functions of these 2 TFs during development, and the generated dataset can therefore be used to extrapolate informative results regarding the genomic landscape regulated by Sox32 and Mixl1 during early development.

I hope that my updated network will provide a useful framework in moving towards a greater understanding of the complex processes controlling early mesendoderm development and provide a link for the molecular changes occurring after gastrulation and the formation of later endodermal derivatives. To the best of my knowledge, this study provides the first in-depth integration of multiple functional genomics data to further decipher mechanisms of gene regulation during endoderm formation in zebrafish; this approach comprises integration of RNA sequencing for gene expression levels of downstream targets of important endodermal

TFs (*sox32* and *mixl1*), with other functional genomic assays such as ChIP-seq to decipher the basic regulatory control exerted by these TFs. This provides a unique and rich set of information that expands our understanding of the impact of interactions instrumental in the coordination of gene expression during mesendoderm specification and endoderm commitment. The systematic integration of my findings with the volume of data produced on zebrafish development in the last decades, including spatial and temporal expression pattern information extracted from ZFIN, significantly expands the endodermal GRN built in 2009 by Chan et al. and the niche mesendodermal GRN assembled by Nelson et al. in 2017. As such, these new findings could ultimately offer valuable knowledge to the broader scientific community for reprogramming stem cells along endodermal cell lineages. This combined information results in an enhanced version of the currently known network architecture governing early endoderm development.

Chapter 7 – Summary of research results

Cell behaviour and tissue development is strictly regulated by a hierarchy of timely and spatially genetic regulatory events. Specific factors have been implicated in the lineage specification of cells of endodermal and mesodermal cell fates in zebrafish. This process is highly regulated and orchestrated by the induction of genes that first separate mesendodermal progenitors into two distinct lineages: endoderm and mesoderm. Later, different subsets of genes are required for cells to transform from endodermal precursors into organs of endodermal lineage for example pancreas, liver or pharyngeal cells, but the mechanisms behind these cell fate decisions are still elusive. To date, little is known about how these general endodermal factors guide the specification of endodermal identity in a spatiotemporal manner; it was the goal of my PhD to contribute to the understanding of this phenomenon and generate a more comprehensive endodermal GRN in zebrafish.

I sought to elucidate the GRN underpinning endoderm development in zebrafish because little is known regarding the specification of this germ layer relative to that of ectoderm and mesoderm. I did so by first determining where two key endodermal TFs, Sox32 and Mixl1, bind in the zebrafish genome via ChIP-exo and then defining how these TFs work. I complemented the information I had garnered from the ChIP with RNA-seq datasets I generated for two endodermal mutant lines (*sox32*$^{-/-}$ and *mixl1*$^{-/-}$) and FACS sorting GFP^{+} cells expressing GFP under the control of the endodermal *sox17* promoter. Together, these data allowed me to interrogate how the transcriptome was affected when the endodermal GRN was perturbed during gastrulation, thus shedding light on the interactions of TFs taking part in it. I uncovered a remarkable number of DEGs that were uniquely present in the mutant lines and several new markers of endoderm in the GFP^{+} cell population. I then proceeded to integrate my data with existing published data sets (ZFIN, RNA-seq time series, scRNA-seq) to update the GRN underpinning endodermal fate during zebrafish development and provide a more integrated endodermal GRN architecture. For the first time, I have bridged the gap between molecular events that occur during gastrulation and the activity of genes orchestrating organogenesis later on during development, by characterising multiple cohorts of genes associated with heart, pharyngeal arch, liver and pancreas development. Hence, my *sox32* dataset at 9.00 hpf helps to explain the sequential activation of genes upon the completion of gastrulation, and links Sox32 to the activation of these organ specific genes, highlighting the

importance of Sox32 in coordinating gene expression between the gastrula stage and organogenesis.

Another line of investigation I pursued as part of this thesis in was to characterise the *sox17*:*GFP* transgenic line, in particular, investigating the specificity of *sox17* promoter activity in controlling the expression of GFP. I wanted to address this question as this is a readily available and highly utilised transgenic line, and although the promoter was correctly turned on in solely endodermal cells in some embryos, it was apparent to me that GFP expression was broader in other embryos and occurred in varying cell types, including those not of endodermal lineage. I believe that understanding the regulation of the promoter in the *sox17*:*GFP* line is important both to correctly characterise this line, as well as in the context of the expanding field of transcriptional adaptation.

7.1 Future directions

As referred to in the discussion sections of previous chapters, some lines of investigation were not completed due to time limitations, however these questions warrant further study.

I generated ChIP-exo data for two endodermal TFs that have allowed us to add new information to the existing endodermal GRN. Although the antibodies used in the ChIP-exo were not sufficiently specific, and both the anti-Sox32 and anti-Mixl1 antibodies pulled down other family members, I showed that even under these circumstances, information can be obtained from analysing these datasets. However, with a more specialised bioinformatics approach, a more integrative conclusion could have been derived from these data. In addition, my results highlight the importance of introducing evolutionary analysis of TF family members when testing for antibody specificity, as only validating against the most closely related family member might not be enough to ensure antibody specificity.

In the zebrafish field, most antibodies are not ChIP-grade and clear standards for antibody validation, as adopted in more advanced consortiums such as the human ENCODE guidelines (Landt et al., 2012) for histones and TF ChIP-seq, are lacking. Strict evaluations are needed when dealing with ChIP-related experiments and it is important to critically assess and compare the quality and reliability of produced datasets. To complete the zebrafish endodermal GRN, a genome-wide search for *sox32* and *sox17* target genes and other endoderm specific TFs will be essential; this analysis will give us insight into the target sites for these

TFs and will increase our understanding of how these TFs regulate endoderm formation. To circumvent the absence of target specific antibodies, a tagged protein strategy could be adopted. Epitope-tagged proteins can be overexpressed (mRNA injection in the embryos) (Xu et al., 2012) or, preferably, integrated in the genome (CRISPR/Cas9 technology to genetically encode protein tagging) and then ChIP performed using a tag specific antibody (Zhang et al., 2008; Savic et al., 2015).

Recently, biotinylating of a tagged TF followed by ChIP-seq, have been successfully applied to zebrafish embryos to reveal bidirectionally transcribed neural crest *cis*-regulatory modules and identify the role of FoxD3 in regulating neural crest migration and differentiation genes (Trinh et al., 2017; Lukoseviciute et al., 2018). Use of biotinylated TFs (He and Pu, 2010; Mazzoni et al., 2011; Matsuda et al., 2017) to study protein-DNA interaction circumvent the problem of protein specific antibodies, which are more prone to crossreact with family members of the protein of interest. Efficient isolation of biotinylated TF-DNA complexes is possible due to high affinity of biotin for streptavidin (Kd = 10^{-15} M) which also allows harsher washing conditions hence decreasing background contamination from spurious binding while enriching for the biotinylated target. The unique profiling toolkit of the aforementioned technique could potentially allow identification of the whole spectrum of genes under the control of *sox32* and *sox17,* without the problem of antibody unspecificity that I encountered. The protocol could help create a model of how the *cis*-regulatory elements logically regulate endoderm development in zebrafish embryos.

As discussed earlier, I generated transcriptomic data from two endodermal mutants whose analyses shed light on the regulatory network downstream of *mixl1* and *sox32*. Differential gene expression analysis identified both known and unknown marker genes; these findings were in agreement with previous *in situ* hybridization expression data and the same technique was used to confirm the new markers. An important conclusion of my analysis was that *sox32* contributes to the process of cell migration and cell adhesion, through regulating gene expression of genes such as *cdh6, pcdh8* and *prex1*. However, it is still unclear whether any influence on cellular migration is derived from a change in migration pattern of the cells expressing such genes, or if the cells expressing them are involved in a paracrine signalling system, affecting not only themselves but also those cells in close proximity. Importantly, I was able to identify some novel downstream genes that mediate Sox32 function in endoderm development such as *txn, met, tiam1a* and *flrt3*; further characterisation of these candidates,

especially with regards to cell migration, is a priority for future work and will build on expanding and exciting research of gastrulation movement (Pézeron et al., 2008; Giger and David, 2017; Liu et al., 2018).

Similarly, the RNA-seq data sets I generated have provided significant new insights into the zebrafish endodermal GRN, but by no means has all the information been extracted and these data sets could be mined further and/or fully integrated. The next steps will be to systematically dissect more genes, whilst in parallel, start to develop a mathematical model to describe the interaction of the genes. This model might, for example, provide a better understanding of how the Nodal/Fgf signalling gradients interact with mesendodermal TFs in the GRN to create an endoderm specification pathway (Poulain et al., 2006; Liu et al., 2018; van Boxtel et al., 2018; Vopalensky et al., 2018).

Further understanding of the regulatory network could also be gained by computational analysis of the promoters of genes identified as differentially regulated in the *sox32*$^{-/-}$ and the *mixl1*$^{-/-}$ mutants. This could yield information about conserved TF binding motifs for Sox32 and Mixl1, or even conserved motifs that can be mapped to additional TFs. Expanding on this idea, the whole 5' flanking regions of genes found to be differentially regulated in both mutants could be scanned for binding motifs recognised by TFs within the known endodermal signalling cascade. Based on the assumption that coexpressed genes that share similar expression patterns over multiple conditions are likely to have similar overrepresented *cis*-elements in their promoter regions (Veerla and Höglund, 2006; Sanchita and Sharma, 2015; Long et al., 2016; van der Graaf et al., 2017; Niwa, 2018); this analysis has the potential of indicating direct regulatory interactions between TFs within the endodermal signalling cascade. Concerning the promoter analysis, we can ask specifically if there are any common motifs underlying the structure of the identified promoters, and whether these features of the promoter architecture can help us to explain how the combinatorial roles of TFs evolve in orchestrating spatiotemporally precise gene expression programs during endoderm development. By combining my transcriptomic data with approaches that characterise the TSS landscape and core promoter sequence features, perhaps we could elucidate further how active genes are switched between the 'on/off position' in a defined manner in this pathway.

Another interesting approach that might help to further elucidate the zebrafish endodermal GRN is to start to integrate information from developmental dynamics of the 3D genome; Hi-C maps have been generated for zebrafish embryos at different time points in development

(Kaaij et al., 2018). Analysis of *Drosophila* development has highlighted the fundamental role of chromatin architecture in regulating gene expression and has uncovered drastic changes in chromatin conformation associated with zygotic genome activation (Hug et al., 2017). Integrating zebrafish Hi-C map information, particularly chromatin organisation and analyses of spatially coexpressed genes within the same topologically associated domains (TADs) (Kaaij et al., 2018), together with the multiple data sets I have generated, could help shed light on the *cis*-regulatory dynamics of endodermal lineage commitment.

All in all, further work is required, but this study represents the first updated review of endoderm development in zebrafish since 2009. Understanding how the complex GRN architecture that contributes to the specification of the germ layers *in vivo* is regulated, is a critical unanswered question in both developmental and evolutionary biology.

The other line of investigation in this research was that of the regulation of gene expression in the *sox17:GFP* transgenic line, which I showed to be directly dependent on *sox17* promoter activity as indicated by GFP expression. In non leaky embryos, *sox17* promotor activity drove GFP expression only in endodermal cells, whereas in leaky embryos, GFP expression occurred in cells of the embryo where the *sox17* promoter should not be active. Due to time constraints, I was unable to conclusively identify the precise consequences of the observed genotypes, but I was able to i) establish a regulatory relationship between the ectopic expression of GFP and resulting changes in transcriptional expression of some genes and ii) establish that there is a compensatory mechanism in place that allows the aberrant transcriptional profile to revert to its native state by 24 hpf, leading to an apparently healthy phenotype.

Genetic compensation by transcriptional modulation of related genes has been described in numerous systems, for example, activation of a compensatory network can act as a buffering mechanism against deleterious mutations (Rossi et al., 2015; El-Brolosy and Stainier, 2017; El-Brolosy et al., 2018). This supports the general idea that there are mechanisms in place to protect a process as complex and fundamental as development from interference, at least to some degree. I am proposing a mechanistic model to explain the unspecific GFP expression in leaky embryos, mainly based on the idea that the integration of the construct has either interrupted a regulatory region important in the control of the *sox17* promoter activity, or that the integrated transgene has undergone some random mutagenesis which either directly interferes with the regulation of the *sox17* promoter activity or interferes with the

activity/expression of an unknown factor X that is involved in the regulation of *sox17* promoter activity.

Recent examples from the literature have shown that even single nucleotide polymorphisms in regulatory regions of a CRM can lead to significant phenotypic variation (Gerke et al., 2009; Pai et al., 2015). Individual polymorphisms in bound sequence motifs can not only influence TF binding and act as a powerful evolutionary driving force but it has become increasingly apparent that polymorphisms and mutations in CRMs are likely to account for significant differences of phenotype variation, resulting in birth defects and chronic diseases (Epstein, 2009; Jones and Swallow, 2011; Deplancke et al., 2016).

Different facets of the transcriptional activation process can be disrupted by a *cis*-acting mutation (Maston et al., 2006), for example, a binding site that had been lost during evolutionary selection of the genome can be restored. Disruption of key TF binding sites could also be explained by a mutation within the TF binding region that i) increases the affinity of normal TFs that act in this CRM, ii) decreases the affinity of normal TFs that act in this CRM or iii) introduces a new binding site. All of these changes can affect the turnover of TFs and TF competition at the binding sites. Moreover, mutations in the CRM can also interfere with target gene expression through other processes, including affecting secondary outputs such as mRNA splicing, stabilisation, degradation and polyadenylation.

Lastly, mutational changes within any individual *cis*-regulatory module are sensitively dependent on the biological context and can affect expression in some tissues without affecting expression in others. For example, Birnbaum et al. (2014) demonstrated how enhancers with distinct mutation profiles differ in different cell types and tissues. Therefore, it can be speculated that if this same scenario were to apply in the *sox17:GFP* line, the factors that increase endodermal GFP^{+} cells during gastrulation are not active during somitogenesis and that the 'misregulated' cells revert to their original fate at later stages of development. Hence, the 'transgene unit' regulation can differ between tissues and time of activation. Tissue enriched and compartmentalised expression of TFs could be another mechanism adopted by a biological system to deter deleterious mutation to effect elsewhere in the body, and in *sox17:gfp* leaky could explain the absence of phenotype at 24 hpf.

A strategy for testing this hypothesis could be to sequence the transgene region upstream of GFP in both leaky and non leaky embryos to identify putative differences in TF binding

behaviour between the conditions, or by adopting a strategy similar to what is described in (Kruse et al., 2019) with chromatin conformation techniques to identify where the transgene is integrated in the genome. Further work to elucidate the mechanism behind this phenomenon is both needed and warranted.

7.2 Concluding remarks

This thesis reveals new insights into the complex biology of endoderm development in zebrafish. I adopted an 'omics' approach to investigate the global genetic program of this germ layer during gastrulation and identified novel genes that play a role in endoderm biology; the data presented in this thesis contribute to the existing knowledge of how these factors integrate and interact in the endodermal GRN. In doing so, I have generated the most up-to-date GRN representing a more comprehensive map of how a cell in the zebrafish embryo becomes committed to an endodermal fate. In particular, dissection of the pathway downstream of *sox32* and *mixl1* highlighted novel roles for both genes, never previously described, validating that the detailed characterisation of single genes and their interactions can provide new and valuable insights. Understanding how cells in the embryo are first specified as endoderm, and then go on to build endodermal structures, will not only increase our fundamental knowledge of developmental events, it will also help us to better understand defects that are involve in endodermal derived organs such liver disease pathogenesis, cystic fibrosis and pancreas adenocarcinoma and provide valuable information for applications such as forward programming and direct reprogramming in disease modelling (Ang et al., 2018; Yiangou et al., 2018). By using core GRN information to understand the mechanisms that lead to disease, we can potentially develop new approaches to derive endoderm tissue for therapeutic purposes (Cheng et al., 2013; Ober and Grapin-Botton, 2015; Yiangou et al., 2018).

In conclusion, I hope that this work not only presents new opportunities to investigate endoderm development but also challenges the reader to reassess how we study developmental biology and genome regulation. We are now guiding a new generation of young researchers enthusiastic to combine the exponential growth of both experimental and computational approaches to tackle big biological questions.

Appendix 1 – *mixl1* mutant top 300 DEGs

Gene Name	baseMean	log2FoldChange	lfcSE	stat	pvalue	padj	symbol	Emsebl_name
ENSDARG00000103599	6475.09233	-8.1891672	0.24295426	-33.706621	4.62E-249	5.11E-245	NA	ENSDARG00000103599
ENSDARG00000091337	585.00311	-6.2239848	0.21676403	-28.713181	2.61E-181	1.93E-177	NA	ENSDARG00000091337
ENSDARG00000087060	2087.38215	-9.7041843	0.37521493	-25.863001	1.74E-147	9.61E-144	NA	ENSDARG00000087060
ENSDARG00000083400	776.262723	-6.2600027	0.24892186	-25.148465	1.47E-139	6.50E-136	NA	ENSDARG00000083400
ENSDARG00000104159	895.724525	-8.8358362	0.395796	-22.324218	2.15E-110	7.93E-107	NA	ENSDARG00000104159
ENSDARG00000082321	360.650334	-6.0880408	0.29165479	-20.874133	9.20E-97	2.91E-93	NA	ENSDARG00000082321
frrs1b	267.196235	-6.6728167	0.3374878	-19.772024	5.19E-87	1.43E-83	frrs1b	ENSDARG00000077605
ENSDARG00000103219	655.442607	-6.5718735	0.33414078	-19.667978	4.06E-86	9.97E-83	NA	ENSDARG00000103219
esyt1b	779.041871	-5.7498293	0.29481853	-19.502945	1.04E-84	2.29E-81	esyt1b	ENSDARG00000014239
ENSDARG00000098703	2838.79199	-3.8749154	0.19886337	-19.485315	1.46E-84	2.94E-81	NA	ENSDARG00000098703
ENSDARG00000100638	5544.49838	6.31583364	0.32452526	19.4617629	2.32E-84	4.27E-81	NA	ENSDARG00000100638
ENSDARG00000106903	15894.0075	-8.7919781	0.46399001	-18.948636	4.53E-80	7.71E-77	NA	ENSDARG00000106903
ENSDARG00000108400	148.166051	-4.8875753	0.26027002	-18.778864	1.12E-78	1.78E-75	NA	ENSDARG00000108400
ENSDARG00000102311	171.419727	5.09999125	0.28058866	18.1760418	7.99E-74	1.18E-70	NA	ENSDARG00000102311
hcn5	171.342125	6.15631221	0.3451971	17.8341943	3.84E-71	5.30E-68	hcn5	ENSDARG00000077382
ENSDARG00000089201	574.502726	-5.4447493	0.30669866	-17.752765	1.64E-70	2.13E-67	NA	ENSDARG00000089201
uchl5	320.499201	-2.0914912	0.1220331	-17.13872	7.63E-66	9.38E-63	uchl5	ENSDARG00000103404
ENSDARG00000083351	483.482935	5.14967396	0.31549285	16.3226328	6.81E-60	7.93E-57	NA	ENSDARG00000083351
ENSDARG00000036895	520.962641	-2.4687913	0.15152526	-16.292935	1.11E-59	1.23E-56	NA	ENSDARG00000036895
znf1156	117.674764	-5.4730123	0.33945834	-16.122781	1.76E-58	1.86E-55	znf1156	ENSDARG00000098800
ENSDARG00000103879	203.389054	2.83464394	0.18188879	15.5844893	9.28E-55	9.33E-52	NA	ENSDARG00000103879
ENSDARG00000090840	138.800676	-3.900589	0.25052157	-15.569873	1.17E-54	1.10E-51	NA	ENSDARG00000090840
zgc:63694	1388.76433	-3.6050831	0.23156768	-15.568161	1.20E-54	1.10E-51	zgc:63694	ENSDARG00000100227
npas2	158.044596	5.74729087	0.37026451	15.522122	2.46E-54	2.17E-51	npas2	ENSDARG00000016536
prkcq	326.711024	-6.500023	0.41979887	-15.48366	4.47E-54	3.81E-51	prkcq	ENSDARG00000034173
ENSDARG00000102771	99.4704641	-3.9623507	0.26697713	-14.841536	7.89E-50	6.47E-47	NA	ENSDARG00000102771
ENSDARG00000104670	261.161064	-4.1781286	0.28447147	-14.687338	7.77E-49	6.14E-46	NA	ENSDARG00000104670
slc7a2	316.676177	-6.788549	0.46239635	-14.681234	8.50E-49	6.49E-46	slc7a2	ENSDARG00000037097

ENSDARG00000095993	113.000563	-3.0933572	0.21414987	-14.444824	2.70E-47	1.99E-44	NA	ENSDARG00000095993
ENSDARG00000092475	109.84089	-3.7345206	0.2599672	-14.365353	8.54E-47	6.09E-44	NA	ENSDARG00000092475
ENSDARG00000103680	110.502055	-4.6552189	0.33061045	-14.080677	4.99E-45	3.45E-42	NA	ENSDARG00000103680
zgc:113208	141.701719	2.67920486	0.19091621	14.0334066	9.74E-45	6.53E-42	zgc:113208	ENSDARG00000035610
ENSDARG00000099400	74.4014855	-4.4299813	0.31643682	-13.999576	1.57E-44	1.02E-41	NA	ENSDARG00000099400
znf1147	108.126624	-4.7639969	0.34562167	-13.783849	3.19E-43	2.01E-40	znf1147	ENSDARG00000100652
ENSDARG00000101103	123.491194	-4.8278727	0.35098449	-13.755231	4.74E-43	2.91E-40	NA	ENSDARG00000101103
ENSDARG00000098139	71.882724	4.28092294	0.31480951	13.5984551	4.09E-42	2.45E-39	NA	ENSDARG00000098139
si:dkeyp-80c12.7	111.885868	7.24942343	0.53882339	13.4541737	2.91E-41	1.69E-38	si:dkeyp-80c12.7	ENSDARG00000052468
si:dkey-23a13.17	2220.73119	-2.3473545	0.17504605	-13.409926	5.29E-41	3.00E-38	si:dkey-23a13.17	ENSDARG00000105328
znf1137	151.874306	-3.8633184	0.28972013	-13.334656	1.46E-40	8.05E-38	znf1137	ENSDARG00000086296
ENSDARG00000100201	79.7178451	-3.428338	0.25893661	-13.240067	5.15E-40	2.78E-37	NA	ENSDARG00000100201
ENSDARG00000100213	124.644553	-6.1878073	0.47406531	-13.052647	6.14E-39	3.23E-36	NA	ENSDARG00000100213
mri1	303.583732	-2.5495332	0.1962995	-12.987976	1.43E-38	7.36E-36	mri1	ENSDARG00000075754
ENSDARG00000098170	114.975025	-4.1095791	0.31670403	-12.976087	1.67E-38	8.41E-36	NA	ENSDARG00000098170
ENSDARG00000101500	20157.9337	1.64630157	0.12741895	12.9203828	3.45E-38	1.70E-35	NA	ENSDARG00000101500
ENSDARG00000098192	109.825766	2.88651196	0.22498904	12.8295672	1.12E-37	5.39E-35	NA	ENSDARG00000098192
ENSDARG00000086668	224.152571	-2.3235324	0.18118287	-12.824239	1.20E-37	5.65E-35	NA	ENSDARG00000086668
col12a1a	293.084644	-7.2998133	0.58066726	-12.571422	3.03E-36	1.40E-33	col12a1a	ENSDARG00000078322
znf1068	139.600491	-2.8878603	0.23096983	-12.503193	7.17E-36	3.24E-33	znf1068	ENSDARG00000078728
znf1072	86.2426227	-5.1271774	0.41554886	-12.338326	5.63E-35	2.49E-32	znf1072	ENSDARG00000096226
usp48	219.08156	-3.2724247	0.26544018	-12.328294	6.38E-35	2.77E-32	usp48	ENSDARG00000090301
ENSDARG00000101489	2713.7354	7.27745665	0.59147695	12.3038719	8.63E-35	3.67E-32	NA	ENSDARG00000101489
ENSDARG00000096152	117.984752	-2.6983844	0.219905	-12.270682	1.30E-34	5.43E-32	NA	ENSDARG00000096152
ENSDARG00000096029	185.119527	-4.2414364	0.34828143	-12.178187	4.06E-34	1.66E-31	NA	ENSDARG00000096029
slc5a2	76.012322	5.7528676	0.4725855	12.1731784	4.32E-34	1.74E-31	slc5a2	ENSDARG00000100919
ENSDARG00000103796	79.6622976	-5.3821356	0.44490558	-12.097254	1.09E-33	4.31E-31	NA	ENSDARG00000103796
nek7	304.783889	1.71616675	0.14205737	12.0808007	1.33E-33	5.18E-31	nek7	ENSDARG00000056966
si:ch211-257p13.3	122.98609	-4.2910769	0.35538621	-12.074405	1.44E-33	5.50E-31	si:ch211-257p13.3	ENSDARG00000053431
pcsk1	100.792314	6.458511	0.53541412	12.062646	1.66E-33	6.24E-31	pcsk1	ENSDARG00000002600

ENSDARG00000101544	160.389802	-2.6980531	0.22376091	-12.05775	1.77E-33	6.51E-31	NA	ENSDARG00000101544
ENSDARG00000105024	426.814739	-2.2049993	0.1832022	-12.035878	2.30E-33	8.35E-31	NA	ENSDARG00000105024
ENSDARG00000081574	963.742372	3.97089775	0.33070182	12.0074868	3.25E-33	1.16E-30	NA	ENSDARG00000081574
ENSDARG00000088847	82.9773414	-2.9346909	0.2454624	-11.955766	6.06E-33	2.13E-30	NA	ENSDARG00000088847
xylb	121.615366	-2.0450348	0.17379612	-11.766861	5.78E-32	2.00E-29	xylb	ENSDARG00000043260
mtmr11	87.916741	4.41983825	0.3771716	11.7183751	1.03E-31	3.49E-29	mtmr11	ENSDARG00000069755
c7b	144.876911	-5.5262017	0.47308102	-11.681301	1.59E-31	5.32E-29	c7b	ENSDARG00000057121
ENSDARG00000102439	91.2831376	-3.129766	0.26802661	-11.677072	1.67E-31	5.51E-29	NA	ENSDARG00000102439
ENSDARG00000093713	289.474844	-2.223563	0.191235	-11.627385	2.99E-31	9.73E-29	NA	ENSDARG00000093713
ENSDARG00000074085	126.439686	-5.8020724	0.50368347	-11.519283	1.05E-30	3.38E-28	NA	ENSDARG00000074085
ENSDARG00000108059	1220.50329	-8.0565502	0.69988148	-11.511307	1.16E-30	3.66E-28	NA	ENSDARG00000108059
ggact.2	75.9076067	-5.4589319	0.47532644	-11.484595	1.58E-30	4.91E-28	ggact.2	ENSDARG00000038248
acy3.2	366.226425	-1.382106	0.12041099	-11.478238	1.70E-30	5.21E-28	acy3.2	ENSDARG00000005525
ENSDARG00000099697	279.126828	-1.9710163	0.17197947	-11.460765	2.08E-30	6.29E-28	NA	ENSDARG00000099697
si:dkey-201g16.1	53.0576801	-5.2545939	0.46007452	-11.42118	3.28E-30	9.80E-28	si:dkey-201g16.1	ENSDARG00000101292
ENSDARG00000096026	63.1381622	-4.8889909	0.43305136	-11.289633	1.48E-29	4.35E-27	NA	ENSDARG00000096026
si:dkey-271j15.3	97.7046801	-2.6431774	0.23488728	-11.252961	2.24E-29	6.52E-27	si:dkey-271j15.3	ENSDARG00000091627
dusp27	78.9334296	4.25247628	0.37887625	11.2239188	3.11E-29	8.94E-27	dusp27	ENSDARG00000099889
si:ch211-39f2.3	62.7906833	4.08328445	0.36413013	11.2138056	3.49E-29	9.89E-27	si:ch211-39f2.3	ENSDARG00000031588
uhrf1bp1	191.414817	2.53252554	0.22683693	11.1645205	6.08E-29	1.70E-26	uhrf1bp1	ENSDARG00000077011
ENSDARG00000100723	46.1308034	-3.5737257	0.32182772	-11.104468	1.19E-28	3.30E-26	NA	ENSDARG00000100723
ENSDARG00000100574	2970.83706	3.24549064	0.29249442	11.0959063	1.31E-28	3.59E-26	NA	ENSDARG00000100574
ENSDARG00000102752	74.5666322	-4.7806335	0.43240123	-11.056013	2.05E-28	5.53E-26	NA	ENSDARG00000102752
ENSDARG00000104745	92.2221644	-2.7821268	0.25173713	-11.051715	2.15E-28	5.73E-26	NA	ENSDARG00000104745
si:ch211-226o13.2	130.421158	-2.5061002	0.22745025	-11.018235	3.12E-28	8.22E-26	si:ch211-226o13.2	ENSDARG00000089875
ENSDARG00000106555	232.640186	-6.0788529	0.55210914	-11.010238	3.41E-28	8.88E-26	NA	ENSDARG00000106555
zgc:171679	67.1427956	-4.3727774	0.397923	-10.989004	4.32E-28	1.11E-25	zgc:171679	ENSDARG00000071024
ENSDARG00000092617	126.58765	-2.8248679	0.25729919	-10.978923	4.83E-28	1.23E-25	NA	ENSDARG00000092617
ENSDARG00000098933	3951.16594	-2.1272884	0.1954557	-10.883737	1.38E-27	3.46E-25	NA	ENSDARG00000098933
si:ch1073-159d7.7	1729.97973	-1.3362935	0.12333111	-10.835008	2.35E-27	5.82E-25	si:ch1073-159d7.7	ENSDARG00000074012

ENSDARG00000100276	434.27903	-2.0994999	0.1937828	-10.834295	2.37E-27	5.82E-25	NA	ENSDARG00000100276
ENSDARG00000051762	399.21069	2.09424355	0.19354814	10.8202722	2.76E-27	6.71E-25	NA	ENSDARG00000051762
rgs4	326.155152	-2.2611643	0.20943666	-10.796411	3.58E-27	8.61E-25	rgs4	ENSDARG00000070047
ENSDARG00000104737	58.9855955	-4.3178623	0.40137656	-10.757635	5.46E-27	1.30E-24	NA	ENSDARG00000104737
scube3	75.1780433	3.85017054	0.35956363	10.7078977	9.35E-27	2.20E-24	scube3	ENSDARG00000011490
ENSDARG00000103779	170.966134	-2.8332311	0.26468116	-10.704317	9.71E-27	2.26E-24	NA	ENSDARG00000103779
il23r	224.877916	-3.6812986	0.34534009	-10.659922	1.57E-26	3.61E-24	il23r	ENSDARG00000052158
ENSDARG00000100001	72.9428353	-3.2988884	0.31042427	-10.627031	2.23E-26	5.09E-24	NA	ENSDARG00000100001
znf1155	147.204298	-1.5695035	0.14888704	-10.541573	5.56E-26	1.25E-23	znf1155	ENSDARG00000100192
kcnh5a	34.0269571	4.29944199	0.40861701	10.5219358	6.85E-26	1.53E-23	kcnh5a	ENSDARG00000043220
map3k5	81.6538672	-2.1034004	0.20054004	-10.488681	9.74E-26	2.15E-23	map3k5	ENSDARG00000005416
plscr3b	1545.05148	-1.546899	0.14766189	-10.475953	1.11E-25	2.44E-23	plscr3b	ENSDARG00000069432
cntn5	241.601554	4.11753469	0.39317622	10.4724916	1.16E-25	2.51E-23	cntn5	ENSDARG00000021584
ENSDARG00000101764	2151.76377	-7.2763403	0.69489647	-10.471114	1.17E-25	2.52E-23	NA	ENSDARG00000101764
mmel1	553.739561	4.13585974	0.39694936	10.4191118	2.03E-25	4.31E-23	mmel1	ENSDARG00000105389
pcyt1bb	63.5290662	7.08982712	0.68495146	10.3508461	4.15E-25	8.74E-23	pcyt1bb	ENSDARG00000104207
ENSDARG00000105106	81.9460376	-2.4301961	0.23499628	-10.341424	4.58E-25	9.55E-23	NA	ENSDARG00000105106
si:dkey-237m9.1	41.4418301	-4.2625191	0.41373019	-10.302654	6.85E-25	1.42E-22	si:dkey-237m9.1	ENSDARG00000098281
si:dkeyp-2e4.3	543.4104	-2.5132491	0.24459078	-10.275322	9.10E-25	1.86E-22	si:dkeyp-2e4.3	ENSDARG00000095283
ENSDARG00000071719	85.550825	-1.9444378	0.18933144	-10.270021	9.62E-25	1.95E-22	NA	ENSDARG00000071719
ENSDARG00000103790	50.6653334	-3.5226516	0.34303482	-10.269079	9.71E-25	1.95E-22	NA	ENSDARG00000103790
ENSDARG00000101843	41.5261486	-5.1527867	0.50181871	-10.268224	9.80E-25	1.95E-22	NA	ENSDARG00000101843
si:dkey-205h13.2	60.6961518	5.157725	0.50300754	10.2537727	1.14E-24	2.25E-22	si:dkey-205h13.2	ENSDARG00000089429
rtn1b	87.8766457	2.18615125	0.21327855	10.250216	1.18E-24	2.31E-22	rtn1b	ENSDARG00000021143
ENSDARG00000088802	6476.60007	4.38125898	0.42753518	10.2477157	1.21E-24	2.35E-22	NA	ENSDARG00000088802
abca4a	113.769729	2.6394277	0.25847808	10.211418	1.76E-24	3.39E-22	abca4a	ENSDARG00000057169
ENSDARG00000101973	60.4119229	-2.5562655	0.25077298	-10.193544	2.12E-24	4.04E-22	NA	ENSDARG00000101973
ENSDARG00000095149	173.920425	-2.6624733	0.26141216	-10.184964	2.31E-24	4.38E-22	NA	ENSDARG00000095149
si:dkey-32n7.4	124.876368	-5.8715761	0.57680879	-10.179415	2.45E-24	4.59E-22	si:dkey-32n7.4	ENSDARG00000002956
ENSDARG00000088378	46.0921321	-4.949502	0.48692165	-10.164884	2.84E-24	5.29E-22	NA	ENSDARG00000088378

churc1	152.601502	-2.0323459	0.20005438	-10.158968	3.02E-24	5.57E-22	churc1	ENSDARG00000010831
ENSDARG00000094459	90.9634264	-2.67565	0.26340446	-10.157952	3.05E-24	5.58E-22	NA	ENSDARG00000094459
ENSDARG00000105179	172.956492	2.96523246	0.29199249	10.1551668	3.14E-24	5.70E-22	NA	ENSDARG00000105179
ENSDARG00000098743	32.6426981	-3.972958	0.3915901	-10.145706	3.46E-24	6.23E-22	NA	ENSDARG00000098743
hebp2	376.366837	-1.2384563	0.12240907	-10.117358	4.63E-24	8.26E-22	hebp2	ENSDARG00000042630
ENSDARG00000099494	63.8238416	-2.6120182	0.2582594	-10.113933	4.79E-24	8.48E-22	NA	ENSDARG00000099494
cry5	475.099847	-1.3626322	0.13523797	-10.075811	7.07E-24	1.24E-21	cry5	ENSDARG00000019498
si:dkey-184n3.2	59.0453697	-4.4793134	0.446244	-10.037812	1.04E-23	1.81E-21	si:dkey-184n3.2	ENSDARG00000093518
s100s	70.88822	-6.0597711	0.60414243	-10.030368	1.12E-23	1.94E-21	s100s	ENSDARG00000036773
cep72	102.444284	-1.9447022	0.19402478	-10.022958	1.21E-23	2.07E-21	cep72	ENSDARG00000105258
ENSDARG00000095522	96.372665	-9.3988263	0.93788093	-10.021343	1.23E-23	2.09E-21	NA	ENSDARG00000095522
ENSDARG00000105940	70.7363744	-2.9395702	0.29476132	-9.9727135	2.01E-23	3.39E-21	NA	ENSDARG00000105940
ENSDARG00000094436	57.5780283	9.39944158	0.94368502	9.96035898	2.27E-23	3.81E-21	NA	ENSDARG00000094436
ENSDARG00000102180	89.754978	4.5690386	0.45886371	9.95728895	2.34E-23	3.90E-21	NA	ENSDARG00000102180
ENSDARG00000086062	442.650482	4.32382567	0.43442223	9.95304876	2.45E-23	4.04E-21	NA	ENSDARG00000086062
thumpd3	162.1768	-2.6493996	0.26757606	-9.9014819	4.10E-23	6.72E-21	thumpd3	ENSDARG00000059634
ENSDARG00000107252	80.6971829	-9.719727	0.98267891	-9.8910508	4.55E-23	7.40E-21	NA	ENSDARG00000107252
ENSDARG00000101765	3174.02902	-5.8874237	0.59770772	-9.8500045	6.85E-23	1.11E-20	NA	ENSDARG00000101765
ENSDARG00000102619	199.286183	-1.336138	0.13577455	-9.8408574	7.51E-23	1.20E-20	NA	ENSDARG00000102619
ENSDARG00000087070	108.976113	-2.886923	0.29406241	-9.8173821	9.48E-23	1.51E-20	NA	ENSDARG00000087070
ENSDARG00000103250	53.9868287	2.99274765	0.30592454	9.78263347	1.34E-22	2.11E-20	NA	ENSDARG00000103250
cyp4v7	60.7677637	-2.6600593	0.2721903	-9.7727923	1.47E-22	2.31E-20	cyp4v7	ENSDARG00000061585
ENSDARG00000100750	130.844712	-2.8983982	0.29780452	-9.7325527	2.19E-22	3.41E-20	NA	ENSDARG00000100750
pkp2	64.7974282	2.87648864	0.29590895	9.72085721	2.46E-22	3.80E-20	pkp2	ENSDARG00000023026
ENSDARG00000103169	42.3088596	-4.9194445	0.50757298	-9.6920931	3.26E-22	5.00E-20	NA	ENSDARG00000103169
ENSDARG00000101310	32.3713826	-4.7860006	0.4963873	-9.641666	5.33E-22	8.13E-20	NA	ENSDARG00000101310
vars	991.053454	-1.047901	0.10871751	-9.6387504	5.49E-22	8.31E-20	vars	ENSDARG00000044575
ENSDARG00000104690	30.4921048	-4.1197207	0.4277135	-9.6319632	5.86E-22	8.82E-20	NA	ENSDARG00000104690
ENSDARG00000109137	168.012857	2.13342144	0.22203215	9.60861514	7.35E-22	1.09E-19	NA	ENSDARG00000109137
LOC108183319	69.8944145	-3.3770867	0.35145133	-9.608974	7.33E-22	1.09E-19	LOC108183319	ENSDARG00000103310
ENSDARG00000106936	81.7757349	-2.2755419	0.23816774	-9.5543669	1.24E-21	1.83E-19	NA	ENSDARG00000106936

hrc	568.003326	-1.3209578	0.1390268	-9.501462	2.07E-21	3.03E-19	hrc	ENSDARG00000045947
ush2a	423.74061	2.67738327	0.28246231	9.47872768	2.57E-21	3.75E-19	ush2a	ENSDARG00000029482
cyp2p6	58.3146877	-2.0651513	0.21803083	-9.4718312	2.75E-21	3.98E-19	cyp2p6	ENSDARG00000042978
chtopa	4942.57139	-0.557105	0.05892086	-9.4551417	3.23E-21	4.63E-19	chtopa	ENSDARG00000057234
cyp2x6	872.31366	1.77326483	0.1875956	9.45259295	3.31E-21	4.72E-19	cyp2x6	ENSDARG00000079653
abcb8	109.676402	-2.3054399	0.24442284	-9.4321786	4.02E-21	5.70E-19	abcb8	ENSDARG00000056672
ENSDARG00000089342	93.231851	-8.2283112	0.87373504	-9.4173987	4.62E-21	6.52E-19	NA	ENSDARG00000089342
akr1a1b	230.400505	-1.672924	0.17821933	-9.3868829	6.18E-21	8.65E-19	akr1a1b	ENSDARG00000052030
mamdc2b	37.8778446	3.66967565	0.39288925	9.34022928	9.61E-21	1.34E-18	mamdc2b	ENSDARG00000073695
si:dkey-110g7.8	70.7739356	-3.7496448	0.40203602	-9.3266388	1.09E-20	1.51E-18	si:dkey-110g7.8	ENSDARG00000074773
LOC100536867	101.83951	-2.014275	0.21604622	-9.3233525	1.13E-20	1.55E-18	LOC100536867	ENSDARG00000102561
ENSDARG00000102347	50.3944792	-4.308315	0.4638119	-9.2889274	1.56E-20	2.12E-18	NA	ENSDARG00000102347
si:ch211-110e21.4	290.199994	-2.1282394	0.22911899	-9.2887955	1.56E-20	2.12E-18	si:ch211-110e21.4	ENSDARG00000076807
ENSDARG00000099030	40.6150581	-3.696109	0.39817171	-9.2827012	1.65E-20	2.23E-18	NA	ENSDARG00000099030
ENSDARG00000080902	113.987805	-4.5862743	0.49412385	-9.281629	1.67E-20	2.24E-18	NA	ENSDARG00000080902
rbm45	1743.14352	1.84788001	0.19928436	9.27257927	1.82E-20	2.42E-18	rbm45	ENSDARG00000063731
si:ch211-276i12.4	103.274204	1.85409483	0.20051845	9.24650493	2.32E-20	3.07E-18	si:ch211-276i12.4	ENSDARG00000100956
cart4	186.465652	-3.4567771	0.37411291	-9.23993	2.47E-20	3.25E-18	cart4	ENSDARG00000070142
ENSDARG00000104666	44.842899	-3.370134	0.36518992	-9.2284421	2.75E-20	3.59E-18	NA	ENSDARG00000104666
ENSDARG00000096166	33.7625916	-4.0807031	0.44310267	-9.209385	3.28E-20	4.27E-18	NA	ENSDARG00000096166
ENSDARG00000094597	246.649378	-3.7807508	0.4111045	-9.1965686	3.70E-20	4.78E-18	NA	ENSDARG00000094597
agbl1	121.681437	1.49083313	0.16222738	9.18977516	3.94E-20	5.06E-18	agbl1	ENSDARG00000104384
ENSDARG00000097091	27.2432169	-4.1497601	0.45299906	-9.1606372	5.16E-20	6.60E-18	NA	ENSDARG00000097091
si:ch211-108d22.2	408.583289	-2.5739333	0.28157898	-9.1410703	6.18E-20	7.86E-18	si:ch211-108d22.2	ENSDARG00000097615
ablim1b	68.1167886	-7.348598	0.80483802	-9.1305305	6.82E-20	8.62E-18	ablim1b	ENSDARG00000045064
ENSDARG00000106172	133.302808	-3.8051558	0.41872733	-9.0874312	1.01E-19	1.27E-17	NA	ENSDARG00000106172
ENSDARG00000104423	49.1031489	-6.2603681	0.69133376	-9.0554931	1.36E-19	1.70E-17	NA	ENSDARG00000104423
scp2a	150.523076	1.44779028	0.16011024	9.04245901	1.53E-19	1.90E-17	scp2a	ENSDARG00000012194
ENSDARG00000042391	52.6462058	-6.3982236	0.70784139	-9.039064	1.58E-19	1.95E-17	NA	ENSDARG00000042391
zgc:55621	100.976568	2.64600219	0.29307817	9.02831553	1.74E-19	2.14E-17	zgc:55621	ENSDARG00000068846

si:ch73-368j24.12	1740.54729	-1.4531801	0.16126971	-9.0108684	2.04E-19	2.50E-17	si:ch73-368j24.12	ENSDARG00000105502
ENSDARG00000076222	51.5108487	-2.9405258	0.32689479	-8.9953276	2.36E-19	2.86E-17	NA	ENSDARG00000076222
ENSDARG00000015607	58.1029515	-6.2989872	0.70074149	-8.9890314	2.49E-19	3.01E-17	NA	ENSDARG00000015607
cbfa2t3	34.0153103	6.19883669	0.6907755	8.97373564	2.87E-19	3.45E-17	cbfa2t3	ENSDARG00000079012
tmem59l	402.114592	-1.7844433	0.19900487	-8.9668324	3.05E-19	3.65E-17	tmem59l	ENSDARG00000003655
ENSDARG00000103172	49.1653783	-4.3424343	0.4857728	-8.9392289	3.92E-19	4.66E-17	NA	ENSDARG00000103172
ENSDARG00000094308	80.057599	-2.7657599	0.31089534	-8.8961123	5.78E-19	6.84E-17	NA	ENSDARG00000094308
ENSDARG00000101426	82.1112739	-2.2318424	0.25135298	-8.8793155	6.73E-19	7.92E-17	NA	ENSDARG00000101426
ENSDARG00000101962	41.4030134	-2.9329243	0.33221181	-8.8284769	1.06E-18	1.24E-16	NA	ENSDARG00000101962
phf11	47.7180757	-2.5971933	0.29430461	-8.8248475	1.10E-18	1.28E-16	phf11	ENSDARG00000021677
si:ch73-30l9.1	63.6350976	-8.2204985	0.93226954	-8.8177272	1.17E-18	1.35E-16	si:ch73-30l9.1	ENSDARG00000095891
ENSDARG00000090400	32.1577252	2.89401958	0.32833819	8.81414238	1.21E-18	1.39E-16	NA	ENSDARG00000090400
ENSDARG00000107478	88.9275884	-2.9074008	0.3301026	-8.8075672	1.28E-18	1.47E-16	NA	ENSDARG00000107478
ENSDARG00000092760	70.2599266	2.43706797	0.27679278	8.80466599	1.31E-18	1.50E-16	NA	ENSDARG00000092760
ENSDARG00000103827	81.0440632	-2.3437739	0.26627146	-8.802197	1.34E-18	1.52E-16	NA	ENSDARG00000103827
ENSDARG00000104432	193.350853	3.32516636	0.37817582	8.79264676	1.46E-18	1.65E-16	NA	ENSDARG00000104432
ENSDARG00000100865	54.4027658	-6.031065	0.68740099	-8.7737218	1.73E-18	1.94E-16	NA	ENSDARG00000100865
si:dkey-27p18.3	71.5049374	-8.390238	0.95656103	-8.7712522	1.77E-18	1.97E-16	si:dkey-27p18.3	ENSDARG00000092228
si:ch211-125e6.8	28.7138287	-4.3395525	0.49542025	-8.759336	1.96E-18	2.18E-16	si:ch211-125e6.8	ENSDARG00000094518
ENSDARG00000106481	32.1787691	-5.4573834	0.6240844	-8.7446241	2.24E-18	2.47E-16	NA	ENSDARG00000106481
gc2	27.9755794	-4.5432905	0.52121526	-8.7167258	2.86E-18	3.15E-16	gc2	ENSDARG00000018329
si:ch73-44m9.1	52.5555262	-6.1604953	0.70755701	-8.7067123	3.13E-18	3.43E-16	si:ch73-44m9.1	ENSDARG00000077115
zgc:171566	913.692449	-1.3403658	0.15395752	-8.7060759	3.15E-18	3.43E-16	zgc:171566	ENSDARG00000105118
si:dkey-17o15.2	126.121506	-1.9832026	0.22788569	-8.7026202	3.24E-18	3.52E-16	si:dkey-17o15.2	ENSDARG00000103956
ENSDARG00000104917	42.8359282	4.01918694	0.4630672	8.67948967	3.98E-18	4.29E-16	NA	ENSDARG00000104917
LOC101882012	52.4783888	-9.0998913	1.04904337	-8.6744663	4.15E-18	4.46E-16	LOC101882012	ENSDARG00000099419
myo7ab	44.3588552	-3.5443334	0.4087767	-8.6705856	4.30E-18	4.59E-16	myo7ab	ENSDARG00000044632
ppfia3	167.133312	2.00198107	0.23104073	8.66505695	4.51E-18	4.80E-16	ppfia3	ENSDARG00000077053
ENSDARG00000090006	40.6487475	3.41672506	0.39444305	8.66215058	4.63E-18	4.90E-16	NA	ENSDARG00000090006
ENSDARG00000092583	137.874124	-1.8620666	0.21555338	-8.6385405	5.69E-18	6.00E-16	NA	ENSDARG00000092583

zgc:171242	30.7243607	5.05469342	0.58591469	8.62701259	6.30E-18	6.60E-16	zgc:171242	ENSDARG00000078551
ENSDARG00000093573	31.2215419	3.93600704	0.45705739	8.61162545	7.20E-18	7.52E-16	NA	ENSDARG00000093573
ENSDARG00000105839	32.9579868	-4.7634212	0.55345907	-8.6066369	7.52E-18	7.81E-16	NA	ENSDARG00000105839
ENSDARG00000102282	56.2082911	-2.5525121	0.29683755	-8.5990202	8.04E-18	8.31E-16	NA	ENSDARG00000102282
si:dkeyp-93d12.1	1460.63809	-1.4266786	0.1659688	-8.5960654	8.25E-18	8.49E-16	si:dkeyp-93d12.1	ENSDARG00000014039
si:ch211-214b16.2	29.5257569	-8.2658359	0.96235627	-8.5891641	8.76E-18	8.97E-16	si:ch211-214b16.2	ENSDARG00000102593
ENSDARG00000104887	54.6391051	-3.2918403	0.38358406	-8.5817966	9.34E-18	9.52E-16	NA	ENSDARG00000104887
ENSDARG00000100468	460.903922	-1.112297	0.13043283	-8.5277381	1.49E-17	1.51E-15	NA	ENSDARG00000100468
ENSDARG00000105930	134.859923	-8.3178809	0.97804319	-8.5046151	1.82E-17	1.84E-15	NA	ENSDARG00000105930
itga10	87.5733832	5.97684366	0.70522438	8.4750951	2.35E-17	2.36E-15	itga10	ENSDARG00000002507
arhgef3	27.1021806	-3.1875778	0.37643671	-8.4677655	2.50E-17	2.50E-15	arhgef3	ENSDARG00000013834
pik3ap1	33.9519937	-3.1266284	0.36994136	-8.4516866	2.87E-17	2.86E-15	pik3ap1	ENSDARG00000078285
ENSDARG00000106512	50.1771673	-3.6487503	0.43176404	-8.4507972	2.89E-17	2.87E-15	NA	ENSDARG00000106512
rapgef4	89.8927141	2.67022298	0.31610049	8.44738643	2.98E-17	2.93E-15	rapgef4	ENSDARG00000079872
LOC100534830	87.5311204	-6.1408622	0.72694821	-8.447455	2.98E-17	2.93E-15	LOC100534830	ENSDARG00000044355
ENSDARG00000108307	37.4243954	-6.1641473	0.7297863	-8.4465101	3.00E-17	2.94E-15	NA	ENSDARG00000108307
ENSDARG00000102371	211.943881	-2.3786402	0.28190398	-8.4377673	3.23E-17	3.15E-15	NA	ENSDARG00000102371
si:dkey-88e18.8	116.940546	3.80139873	0.45070962	8.43425254	3.33E-17	3.23E-15	si:dkey-88e18.8	ENSDARG00000100256
si:ch1073-390k14.1	78.9048473	-3.762085	0.44713152	-8.4138221	3.97E-17	3.83E-15	si:ch1073-390k14.1	ENSDARG00000088549
ENSDARG00000093478	32.9880265	-5.9723758	0.71027655	-8.4085218	4.15E-17	3.99E-15	NA	ENSDARG00000093478
kirrel3a	69.2829769	-3.3000233	0.3925707	-8.4061885	4.24E-17	4.06E-15	kirrel3a	ENSDARG00000075806
alox5a	141.927949	-5.7051026	0.67943928	-8.3967806	4.59E-17	4.38E-15	alox5a	ENSDARG00000057273
serpinb1l2	38.2690616	-2.719853	0.32431959	-8.3863357	5.02E-17	4.76E-15	serpinb1l2	ENSDARG00000070396
tnfb	151.818189	1.59191892	0.19020771	8.36937131	5.79E-17	5.48E-15	tnfb	ENSDARG00000013598
ENSDARG00000098168	2136.6701	-8.7523844	1.04691097	-8.3601994	6.26E-17	5.89E-15	NA	ENSDARG00000098168
pstpip2	52.4148002	3.44914277	0.41294388	8.35257022	6.68E-17	6.26E-15	pstpip2	ENSDARG00000089569
ENSDARG00000104561	103.829831	3.69452781	0.44307507	8.33837888	7.53E-17	7.03E-15	NA	ENSDARG00000104561
cyb561d1	38.4803468	-3.3296134	0.39953603	-8.3337	7.84E-17	7.28E-15	cyb561d1	ENSDARG00000055295
ENSDARG00000080448	228.622035	-2.7406886	0.32911728	-8.3273921	8.26E-17	7.65E-15	NA	ENSDARG00000080448
si:dkey-169i5.4	29.380658	4.42604419	0.53194921	8.32042625	8.76E-17	8.08E-15	si:dkey-169i5.4	ENSDARG00000077862

ENSDARG00000101817	30.5861994	-3.0428138	0.36580005	-8.3182434	8.93E-17	8.19E-15	NA	ENSDARG00000101817
ENSDARG00000091243	47.9614741	-2.4196123	0.29106346	-8.3130063	9.33E-17	8.53E-15	NA	ENSDARG00000091243
ENSDARG00000096700	61.7679817	-9.3333208	1.12546897	-8.2928282	1.11E-16	1.01E-14	NA	ENSDARG00000096700
ENSDARG00000099357	22.5783909	-7.8836307	0.95511924	-8.2540801	1.53E-16	1.39E-14	NA	ENSDARG00000099357
ap4e1	273.813028	1.02705892	0.12449659	8.24969521	1.59E-16	1.43E-14	ap4e1	ENSDARG00000103684
fstl1b	565.455636	-2.3390456	0.28403555	-8.2350451	1.79E-16	1.61E-14	fstl1b	ENSDARG00000039576
ENSDARG00000098892	137.400011	-2.1101528	0.25636268	-8.2311233	1.85E-16	1.66E-14	NA	ENSDARG00000098892
ENSDARG00000100187	21.2465373	-7.795406	0.94717249	-8.2301863	1.87E-16	1.67E-14	NA	ENSDARG00000100187
ENSDARG00000076252	46.0054993	-2.0850815	0.25352758	-8.224279	1.96E-16	1.74E-14	NA	ENSDARG00000076252
ENSDARG00000077719	84.709179	-4.1203712	0.50190839	-8.2094088	2.22E-16	1.97E-14	NA	ENSDARG00000077719
ENSDARG00000056145	81.3899908	-5.4098028	0.66010822	-8.1953271	2.50E-16	2.20E-14	NA	ENSDARG00000056145
ENSDARG00000098548	47.7354939	6.41281836	0.78278991	8.19225984	2.56E-16	2.25E-14	NA	ENSDARG00000098548
ENSDARG00000105097	3216.93302	-8.5403955	1.04694363	-8.157455	3.42E-16	2.99E-14	NA	ENSDARG00000105097
si:ch211-127b11.1	26.7006182	-8.1272565	0.99970905	-8.1296217	4.31E-16	3.72E-14	si:ch211-127b11.1	ENSDARG00000077090
dync1i2a	1043.25642	-0.6105659	0.07510418	-8.1295864	4.31E-16	3.72E-14	dync1i2a	ENSDARG00000078386
ENSDARG00000098483	3982.35263	3.61401032	0.44453955	8.12978353	4.30E-16	3.72E-14	NA	ENSDARG00000098483
grb14	23.9238595	4.02080284	0.49527054	8.1183969	4.72E-16	4.06E-14	grb14	ENSDARG00000068280
LOC101883645	51.089393	-7.3511253	0.90553001	-8.1180361	4.74E-16	4.06E-14	LOC101883645	ENSDARG00000040640
ENSDARG00000102642	40.0697234	-3.1938559	0.39367282	-8.1129703	4.94E-16	4.22E-14	NA	ENSDARG00000102642
zgc:113176	456.563604	1.38666467	0.17097419	8.1103742	5.05E-16	4.29E-14	zgc:113176	ENSDARG00000040487
ENSDARG00000108793	29.0481629	5.46401644	0.67377832	8.10951657	5.08E-16	4.31E-14	NA	ENSDARG00000108793
ENSDARG00000097611	24.0375303	-7.9682221	0.9831412	-8.1048604	5.28E-16	4.46E-14	NA	ENSDARG00000097611
ENSDARG00000100860	387.356846	-3.3066315	0.40846226	-8.0953169	5.71E-16	4.80E-14	NA	ENSDARG00000100860
ENSDARG00000105055	28.6483586	-8.2231937	1.01582573	-8.0950831	5.72E-16	4.80E-14	NA	ENSDARG00000105055
ENSDARG00000103958	28.4280633	-5.7574254	0.71148962	-8.0920722	5.87E-16	4.88E-14	NA	ENSDARG00000103958
tubb5	827.63435	1.38653437	0.17134481	8.09207096	5.87E-16	4.88E-14	tubb5	ENSDARG00000037997
rnd1a	26.7315299	2.88515951	0.35676345	8.08703776	6.11E-16	5.06E-14	rnd1a	ENSDARG00000030547
ENSDARG00000097522	214.462055	-3.283149	0.40623547	-8.0818867	6.38E-16	5.26E-14	NA	ENSDARG00000097522
cyr61l2	109.377114	-3.306934	0.40957717	-8.0740194	6.80E-16	5.59E-14	cyr61l2	ENSDARG00000099985
csrp3	36.1486441	3.39624129	0.42132156	8.06092453	7.57E-16	6.20E-14	csrp3	ENSDARG00000101706
grip1	191.653268	1.6516931	0.20507197	8.05421183	8.00E-16	6.53E-14	grip1	ENSDARG00000015053

si:ch1073-513e17.1	54.1846076	-4.1991256	0.52229194	-8.0398055	9.00E-16	7.32E-14	si:ch1073-513e17.1	ENSDARG00000086739
LOC559843	210.220582	-3.2901105	0.40933637	-8.0376697	9.16E-16	7.42E-14	LOC559843	ENSDARG00000028784
ENSDARG00000083379	362.426044	3.79436886	0.47224391	8.03476501	9.38E-16	7.57E-14	NA	ENSDARG00000083379
LOC108182861	137.375751	-2.8179212	0.35120978	-8.0234703	1.03E-15	8.27E-14	LOC108182861	ENSDARG00000103947
ndrg1b	236.445558	3.98237411	0.49636878	8.02301487	1.03E-15	8.27E-14	ndrg1b	ENSDARG00000010420
ENSDARG00000100640	28.6944334	-7.6456521	0.95403074	-8.0140522	1.11E-15	8.86E-14	NA	ENSDARG00000100640
ENSDARG00000107931	64.8350755	-2.6850122	0.33508208	-8.0129985	1.12E-15	8.91E-14	NA	ENSDARG00000107931
ENSDARG00000022461	169.771022	1.46917937	0.18337112	8.01205432	1.13E-15	8.94E-14	NA	ENSDARG00000022461
znf1034	29.1153562	-3.2392868	0.40529397	-7.9924376	1.32E-15	1.05E-13	znf1034	ENSDARG00000102994
ENSDARG00000091176	200.451165	-2.1416343	0.26805217	-7.9896178	1.35E-15	1.07E-13	NA	ENSDARG00000091176
mthfr	274.399943	-0.9046547	0.11323563	-7.989135	1.36E-15	1.07E-13	mthfr	ENSDARG00000053087
LOC101885092	49.2874593	-3.7431456	0.46919021	-7.9778852	1.49E-15	1.16E-13	LOC101885092	ENSDARG00000095139
ENSDARG00000100705	12193.2839	1.25473313	0.15731582	7.97588635	1.51E-15	1.18E-13	NA	ENSDARG00000100705
LOC100148466	29.3159341	-2.7683798	0.34726999	-7.9718372	1.56E-15	1.21E-13	LOC100148466	ENSDARG00000092436
spata7	89.050065	-1.5578168	0.19541814	-7.9717104	1.56E-15	1.21E-13	spata7	ENSDARG00000075898
ENSDARG00000104397	30.3428891	-3.5963374	0.45124724	-7.9697716	1.59E-15	1.23E-13	NA	ENSDARG00000104397
ENSDARG00000105144	47.1589842	-3.6731129	0.46131667	-7.9622375	1.69E-15	1.30E-13	NA	ENSDARG00000105144
wtip	44.03914	-3.7059229	0.46546831	-7.9617083	1.70E-15	1.30E-13	wtip	ENSDARG00000103607
khk	197.141817	5.7703673	0.72608015	7.94728696	1.91E-15	1.45E-13	khk	ENSDARG00000029874
si:ch211-152c8.5	752.86074	-3.0845538	0.38838257	-7.9420501	1.99E-15	1.51E-13	si:ch211-152c8.5	ENSDARG00000104576
si:ch211-162i8.2	84.6203494	-1.4777255	0.18687878	-7.9074015	2.63E-15	1.98E-13	si:ch211-162i8.2	ENSDARG00000096041
rab23	163.344277	-1.0665531	0.13488081	-7.9073745	2.63E-15	1.98E-13	rab23	ENSDARG00000004151
ENSDARG00000096215	131.798221	-8.3027698	1.05077925	-7.9015357	2.75E-15	2.07E-13	NA	ENSDARG00000096215
ENSDARG00000077877	71.3102282	-4.4686475	0.56586647	-7.897	2.86E-15	2.14E-13	NA	ENSDARG00000077877
uap1l1	113.548809	-1.9414187	0.24588271	-7.8957106	2.89E-15	2.16E-13	uap1l1	ENSDARG00000013082
LOC108183901	62.7021567	-1.8851183	0.23877171	-7.8950656	2.90E-15	2.16E-13	LOC108183901	ENSDARG00000096189
wu:fl23c11	45.871134	-2.8809097	0.36507541	-7.8912731	2.99E-15	2.22E-13	wu:fl23c11	ENSDARG00000105863
ENSDARG00000103307	53.22129	2.41455102	0.30694087	7.86650219	3.65E-15	2.70E-13	NA	ENSDARG00000103307
si:rp71-1h20.5	323.885998	-1.2662509	0.16127939	-7.8512874	4.12E-15	3.04E-13	si:rp71-1h20.5	ENSDARG00000098176
si:dkey-15h8.17	23.9068686	-3.83029	0.48906173	-7.8319152	4.80E-15	3.53E-13	si:dkey-15h8.17	ENSDARG00000099428

Appendix 2 – *sox32* mutant 5.25 hpf top 300 DEGs

Gene Name	baseMean	log2FoldChange	lfcSE	stat	pvalue	padj	symbol	Emsebl_name
pck2	303.91065	-1.7709335	0.24060512	-7.3603315	1.83E-13	1.59E-09	pck2	ENSDARG00000020956
hkdc1	936.875221	-1.0610209	0.15310738	-6.9299134	4.21E-12	1.83E-08	hkdc1	ENSDARG00000038703
pleca	880.683382	-0.885286	0.13969611	-6.337227	2.34E-10	6.76E-07	pleca	ENSDARG00000062590
sbf2	193.484522	1.63643372	0.27325148	5.98874612	2.11E-09	3.06E-06	sbf2	ENSDARG00000059460
nipsnap1	585.724203	-0.9359816	0.15500112	-6.0385475	1.56E-09	3.06E-06	nipsnap1	ENSDARG00000005320
abcc6b.2	1166.18968	-0.7608231	0.12699344	-5.9910424	2.09E-09	3.06E-06	abcc6b.2	ENSDARG00000094901
abcb5	983.733694	-0.9350938	0.15795978	-5.919822	3.22E-09	3.60E-06	abcb5	ENSDARG00000021787
mxtx2	689.713128	0.88560429	0.14972888	5.91471935	3.32E-09	3.60E-06	mxtx2	ENSDARG00000015906
ENSDARG00000010124	1626.51414	-0.8927388	0.15353743	-5.8144702	6.08E-09	5.86E-06	NA	ENSDARG00000010124
mmp30	830.920253	-0.8243963	0.142492	-5.7855621	7.23E-09	6.27E-06	mmp30	ENSDARG00000045887
erbb3a	518.85309	-0.9339406	0.16322653	-5.7217453	1.05E-08	7.44E-06	erbb3a	ENSDARG00000006202
tfa	379.674216	-1.1231078	0.19679134	-5.7070994	1.15E-08	7.44E-06	tfa	ENSDARG00000016771
apoc1	3417.53994	-0.649517	0.11396007	-5.6995139	1.20E-08	7.44E-06	apoc1	ENSDARG00000092170
nanog	5839.69643	0.63274384	0.11070131	5.71577559	1.09E-08	7.44E-06	nanog	ENSDARG00000075113
notum1a	3033.34228	-0.5825498	0.10336791	-5.6356933	1.74E-08	1.01E-05	notum1a	ENSDARG00000031126
gadd45ga	447.758959	-0.9391578	0.16731404	-5.613144	1.99E-08	1.08E-05	gadd45ga	ENSDARG00000019417
lama1	1882.47425	-0.7064002	0.12631216	-5.5924953	2.24E-08	1.14E-05	lama1	ENSDARG00000102277
cdh6	1360.90801	-0.6511846	0.11917589	-5.4640628	4.65E-08	2.24E-05	cdh6	ENSDARG00000014522
blf	6345.42412	0.57424945	0.1065323	5.39037865	7.03E-08	3.05E-05	blf	ENSDARG00000043126
urahb	385.575376	-0.9580146	0.17754537	-5.3958862	6.82E-08	3.05E-05	urahb	ENSDARG00000089331
zgc:123010	690.758978	0.80006787	0.14984706	5.33922956	9.33E-08	3.85E-05	zgc:123010	ENSDARG00000039082
slc43a2b	2579.3053	-0.5889457	0.11119527	-5.2964996	1.18E-07	4.45E-05	slc43a2b	ENSDARG00000061120
ENSDARG00000099121	275.336077	-1.1524676	0.21753889	-5.2977543	1.17E-07	4.45E-05	NA	ENSDARG00000099121
msgn1	529.855547	-0.8971653	0.17044866	-5.2635512	1.41E-07	5.10E-05	msgn1	ENSDARG00000070546
ttc22	430.734149	-0.9344405	0.17903545	-5.2193044	1.80E-07	6.23E-05	ttc22	ENSDARG00000077836
cahz	172.925307	-1.3271518	0.25646484	-5.1747905	2.28E-07	7.61E-05	cahz	ENSDARG00000011166
fetub	1037.96509	-0.8482215	0.16523926	-5.133293	2.85E-07	9.14E-05	fetub	ENSDARG00000053973
dlc	2343.41114	-0.6211899	0.12383069	-5.0164452	5.26E-07	0.00016298	dlc	ENSDARG00000002336
aldh1a2	542.440835	-0.8198977	0.16472995	-4.977223	6.45E-07	0.00019284	aldh1a2	ENSDARG00000053493

hes6	1627.68212	-0.5886463	0.11848988	-4.9679032	6.77E-07	0.00019426	hes6	ENSDARG00000019335
gtf2a1	6403.06749	0.49899514	0.10054556	4.96287586	6.95E-07	0.00019426	gtf2a1	ENSDARG00000011000
tbxta	8683.58329	-0.4832714	0.09818395	-4.9221014	8.56E-07	0.00023198	tbxta	ENSDARG00000101576
ENSDARG00000093494	845.646134	-0.6728913	0.13735966	-4.8987551	9.64E-07	0.00025339	NA	ENSDARG00000093494
cdx4	381.325331	-0.9526219	0.1952047	-4.8801177	1.06E-06	0.00027036	cdx4	ENSDARG00000036292
s100a1	2882.57869	0.52332566	0.10744924	4.87044546	1.11E-06	0.00027582	s100a1	ENSDARG00000015543
rwdd	575.921527	-0.7674673	0.15794552	-4.8590637	1.18E-06	0.00028404	rwdd	ENSDARG00000068256
camsap3	1018.72245	-0.6540473	0.13555323	-4.825022	1.40E-06	0.00032803	camsap3	ENSDARG00000059475
dspa	5404.25536	-0.4903911	0.10187506	-4.813652	1.48E-06	0.00032945	dspa	ENSDARG00000022309
mtfr1	334.935285	0.95848758	0.19901039	4.81626909	1.46E-06	0.00032945	mtfr1	ENSDARG00000045304
twsg1b	1536.22202	-0.5671128	0.11830058	-4.7938298	1.64E-06	0.00035466	twsg1b	ENSDARG00000103580
hprt1	1069.01175	0.72179226	0.15203636	4.74749747	2.06E-06	0.00043551	hprt1	ENSDARG00000008884
cdt1	17400.9651	0.41734647	0.08830356	4.72627004	2.29E-06	0.00047206	cdt1	ENSDARG00000051854
lcp1	2294.23093	-0.607279	0.12871055	-4.718176	2.38E-06	0.00047581	lcp1	ENSDARG00000023188
mkrn4	7692.03101	0.46550762	0.09872486	4.71520172	2.41E-06	0.00047581	mkrn4	ENSDARG00000028295
kif15	1364.01515	-0.568928	0.12093779	-4.7043033	2.55E-06	0.00049079	kif15	ENSDARG00000012073
ccna1	2959.06381	0.57256533	0.12393676	4.61981835	3.84E-06	0.0007239	ccna1	ENSDARG00000043236
si:dkey-166k12.1	415.283916	-0.7926133	0.17251746	-4.5943949	4.34E-06	0.0008006	si:dkey-166k12.1	ENSDARG00000054690
dld	1302.25248	-0.7062201	0.15473148	-4.5641653	5.01E-06	0.00089979	dld	ENSDARG00000020219
zgc:56676	6567.94729	0.42476732	0.09312548	4.56123623	5.09E-06	0.00089979	zgc:56676	ENSDARG00000029445
ywhae2	4264.29683	-0.5112268	0.11220363	-4.556241	5.21E-06	0.00090302	ywhae2	ENSDARG00000017014
nop2	2474.60026	-0.5438787	0.11980952	-4.5395287	5.64E-06	0.00095846	nop2	ENSDARG00000043304
smg1	7073.90123	-0.4557469	0.10070161	-4.5257166	6.02E-06	0.00098513	smg1	ENSDARG00000054570
spdl1	1628.58536	-0.531555	0.11745487	-4.5256101	6.02E-06	0.00098513	spdl1	ENSDARG00000103996
dpp7	343.890732	-1.0511891	0.23582326	-4.4575293	8.29E-06	0.00130696	dpp7	ENSDARG00000027750
tmem192	728.287063	0.67323708	0.15102971	4.45764668	8.29E-06	0.00130696	tmem192	ENSDARG00000037484
acy3.2	897.548397	-0.6863472	0.15524595	-4.4210312	9.82E-06	0.00149415	acy3.2	ENSDARG00000005525
ENSDARG00000106126	317.112025	-0.8481289	0.19167512	-4.4248253	9.65E-06	0.00149415	NA	ENSDARG00000106126
ENSDARG00000075627	1211.99084	0.54346198	0.12332806	4.40663677	1.05E-05	0.00156939	NA	ENSDARG00000075627
kif11	5292.76264	-0.4206344	0.0957665	-4.3922915	1.12E-05	0.00162991	kif11	ENSDARG00000010948
si:ch73-347e22.8	934.057606	-0.6662542	0.15172949	-4.3910657	1.13E-05	0.00162991	si:ch73-347e22.8	ENSDARG00000103322

mpp7a	486.654205	0.70585355	0.16274548	4.3371622	1.44E-05	0.00205144	mpp7a	ENSDARG00000102470
slc16a3	3368.21108	0.48016652	0.11142477	4.30933366	1.64E-05	0.00228982	slc16a3	ENSDARG00000045051
efnb2a	1339.80326	-0.5468098	0.12723028	-4.2977963	1.73E-05	0.00236249	efnb2a	ENSDARG00000020164
blvrb	1064.02417	0.53971045	0.12564903	4.2953809	1.74E-05	0.00236249	blvrb	ENSDARG00000096829
hmgb3a	5094.69591	-0.4566786	0.10649944	-4.2880842	1.80E-05	0.00240387	hmgb3a	ENSDARG00000056725
rnf144b	904.435815	-0.64682	0.1514793	-4.2700225	1.95E-05	0.00256754	rnf144b	ENSDARG00000024940
pcyt1aa	1243.13732	0.5148855	0.12071979	4.26512913	2.00E-05	0.0025853	pcyt1aa	ENSDARG00000011233
arl6ip5a	1897.08742	-0.4677431	0.1100814	-4.2490656	2.15E-05	0.00273269	arl6ip5a	ENSDARG00000077044
a1cf	193.56568	-1.3459216	0.31697502	-4.246144	2.17E-05	0.00273269	a1cf	ENSDARG00000002968
mcm6	7718.45663	-0.4415086	0.10416273	-4.2386428	2.25E-05	0.00276623	mcm6	ENSDARG00000057683
zswim5	4123.9225	-0.4448711	0.10507476	-4.2338528	2.30E-05	0.00276623	zswim5	ENSDARG00000055900
acsf2	3085.43733	0.4596176	0.10847861	4.23694223	2.27E-05	0.00276623	acsf2	ENSDARG00000061201
smchd1	2015.59868	0.52935463	0.12523366	4.22693566	2.37E-05	0.00281354	smchd1	ENSDARG00000104374
ENSDARG00000106807	193.535965	1.30296687	0.31056684	4.19544749	2.72E-05	0.00303678	NA	ENSDARG00000106807
anxa1b	505.896951	-0.7690666	0.18334144	-4.1947234	2.73E-05	0.00303678	anxa1b	ENSDARG00000100095
slc38a3b	2898.61023	-0.4536193	0.10800901	-4.1998282	2.67E-05	0.00303678	slc38a3b	ENSDARG00000091061
cenpe	2518.33335	-0.4802472	0.11428335	-4.2022499	2.64E-05	0.00303678	cenpe	ENSDARG00000063385
nnr	3189.27421	0.4706791	0.11191235	4.20578319	2.60E-05	0.00303678	nnr	ENSDARG00000058917
si:ch211-66k16.27	233.114797	-1.0618704	0.25352776	-4.1883793	2.81E-05	0.00308338	si:ch211-66k16.27	ENSDARG00000091280
slc4a10a	969.807873	-0.5630635	0.13466045	-4.1813575	2.90E-05	0.00314042	slc4a10a	ENSDARG00000063133
ccne1	9327.09241	0.38792234	0.09313312	4.16524585	3.11E-05	0.00332904	ccne1	ENSDARG00000098622
myzap	169.162841	-1.1574784	0.27881602	-4.1514056	3.30E-05	0.00337048	myzap	ENSDARG00000075017
ENSDARG00000079688	6835.18386	-0.4560845	0.10971945	-4.1568244	3.23E-05	0.00337048	NA	ENSDARG00000079688
lmo7b	1030.10611	0.71484851	0.17218333	4.15167079	3.30E-05	0.00337048	lmo7b	ENSDARG00000053535
numa1	3300.93612	-0.4526806	0.10889916	-4.1568787	3.23E-05	0.00337048	numa1	ENSDARG00000102483
top2a	6977.74317	-0.403933	0.09783189	-4.128848	3.65E-05	0.00367553	top2a	ENSDARG00000024488
fdx1	1152.12382	0.54257441	0.13172916	4.11886345	3.81E-05	0.00379433	fdx1	ENSDARG00000056410
zgc:56585	306.194334	-1.1244887	0.27328886	-4.1146525	3.88E-05	0.00382035	zgc:56585	ENSDARG00000026236
nckap5l	896.305752	-0.5820195	0.14377753	-4.0480559	5.16E-05	0.00498102	nckap5l	ENSDARG00000079148
cyp2aa9	622.736754	0.68377608	0.16892627	4.04777827	5.17E-05	0.00498102	cyp2aa9	ENSDARG00000098890
micall2a	1787.50382	0.57619236	0.14275329	4.03628078	5.43E-05	0.00517391	micall2a	ENSDARG00000102366

fam49ba	1337.79175	0.48562109	0.12086291	4.01794954	5.87E-05	0.0054852	fam49ba	ENSDARG00000020929
smarcad1a	1804.36803	-0.4509011	0.11223636	-4.0174243	5.88E-05	0.0054852	smarcad1a	ENSDARG00000014041
pcdh8	2010.38567	-0.4623303	0.11548688	-4.0033142	6.25E-05	0.00576105	pcdh8	ENSDARG00000006467
her1	1169.48863	-0.5622156	0.14093795	-3.9890999	6.63E-05	0.00594205	her1	ENSDARG00000014722
znf326	805.006585	0.58866	0.14741745	3.99315018	6.52E-05	0.00594205	znf326	ENSDARG00000098348
atf1	3678.45979	0.39450951	0.09909031	3.98131265	6.85E-05	0.00594205	atf1	ENSDARG00000044301
si:dkeyp-13a3.10	592.444027	-0.6963975	0.17475567	-3.9849783	6.75E-05	0.00594205	si:dkeyp-13a3.10	ENSDARG00000092225
plcg1	2811.9653	0.42046811	0.10547417	3.98645578	6.71E-05	0.00594205	plcg1	ENSDARG00000038442
pip4k2ca	2868.6935	0.43289699	0.10871518	3.98193711	6.84E-05	0.00594205	pip4k2ca	ENSDARG00000031020
hmmr	1978.45422	-0.4762738	0.11986326	-3.9734762	7.08E-05	0.00602066	hmmr	ENSDARG00000021794
btf3l4	2082.70094	0.42873848	0.10789873	3.97352662	7.08E-05	0.00602066	btf3l4	ENSDARG00000089681
ENSDARG00000093220	303.376986	0.87578011	0.22099115	3.9629646	7.40E-05	0.00623102	NA	ENSDARG00000093220
mast1b	444.063757	-0.7099477	0.18006363	-3.9427599	8.05E-05	0.00658832	mast1b	ENSDARG00000088789
tarbp2	2305.30491	0.42806676	0.1085457	3.94365465	8.02E-05	0.00658832	tarbp2	ENSDARG00000070471
epas1a	597.856336	-0.6233913	0.15804758	-3.944327	8.00E-05	0.00658832	epas1a	ENSDARG00000008697
lamb1a	3484.06244	-0.4976885	0.12634895	-3.9389997	8.18E-05	0.00662988	lamb1a	ENSDARG00000101209
xkr5b	286.89378	-0.9548424	0.24312105	-3.9274363	8.59E-05	0.00689234	xkr5b	ENSDARG00000097530
krcp	3481.92904	-0.5736427	0.14706577	-3.900586	9.60E-05	0.00736261	krcp	ENSDARG00000040224
cspp1b	598.472582	-0.6609931	0.16939211	-3.9021479	9.53E-05	0.00736261	cspp1b	ENSDARG00000091628
tmsb1	1013.23416	-0.5876794	0.15056592	-3.9031372	9.50E-05	0.00736261	tmsb1	ENSDARG00000104181
im:7160594	1352.01708	0.5602699	0.14363555	3.90063525	9.59E-05	0.00736261	im:7160594	ENSDARG00000070957
si:dkeyp-34c12.1	343.436397	0.85697039	0.21953245	3.90361599	9.48E-05	0.00736261	si:dkeyp-34c12.1	ENSDARG00000071083
sacs	336.098495	0.75293473	0.1935384	3.89036346	0.00010009	0.00761243	sacs	ENSDARG00000091042
celsr1a	947.141814	-0.5542792	0.14292817	-3.8780266	0.00010531	0.00793256	celsr1a	ENSDARG00000069185
entpd1	1253.84846	-0.4909928	0.12667107	-3.8761244	0.00010613	0.00793256	entpd1	ENSDARG00000045066
aktip	1064.66878	0.53222113	0.13751007	3.87041561	0.00010865	0.00805124	aktip	ENSDARG00000026862
tesca	343.117731	-0.7936116	0.20540281	-3.8636844	0.00011169	0.00814007	tesca	ENSDARG00000028346
knl1	1539.40848	-0.4893487	0.12672307	-3.8615601	0.00011267	0.00814007	knl1	ENSDARG00000070239
si:ch211-285f17.1	635.066923	-0.584865	0.15140866	-3.862824	0.00011208	0.00814007	si:ch211-285f17.1	ENSDARG00000059333
mllt1b	683.156052	-0.5881584	0.15262074	-3.8537254	0.00011633	0.00831659	mllt1b	ENSDARG00000031709
si:dkeyp-114g9.1	6552.79221	0.38194262	0.09914737	3.85227186	0.00011703	0.00831659	si:dkeyp-114g9.1	ENSDARG00000069401

pla2g12b	324.099605	-0.7719198	0.20202413	-3.8209289	0.00013295	0.00937135	pla2g12b	ENSDARG00000015662
epb41l5	4359.54217	0.37934741	0.09953848	3.81106281	0.00013837	0.00967478	epb41l5	ENSDARG00000032324
rap1ab	1015.48856	0.52815402	0.13872672	3.80715423	0.00014058	0.00975029	rap1ab	ENSDARG00000087346
kpnb1	7121.75324	-0.3729455	0.0981141	-3.8011411	0.00014403	0.00991072	kpnb1	ENSDARG00000104889
rftn2	335.37817	0.70765227	0.18652174	3.79393989	0.00014828	0.01012244	rftn2	ENSDARG00000056078
rnpc3	794.132348	-0.5293471	0.13969661	-3.789262	0.0001511	0.01023436	rnpc3	ENSDARG00000011247
tent5ba	3001.35636	0.39008789	0.10322742	3.7789173	0.00015751	0.01058625	tent5ba	ENSDARG00000039943
p4ha2	371.818979	-0.7631448	0.20213676	-3.7753884	0.00015976	0.01065466	p4ha2	ENSDARG00000010085
fam113	689.36794	0.61454255	0.16300415	3.77010369	0.00016318	0.01078895	fam113	ENSDARG00000002685
dsc2l	1021.90691	-0.517059	0.13720713	-3.7684555	0.00016426	0.01078895	dsc2l	ENSDARG00000039677
celsr2	350.534773	-0.7118937	0.18905151	-3.7656073	0.00016614	0.01083064	celsr2	ENSDARG00000019726
adcy5	357.703001	0.72813198	0.19376119	3.75788339	0.00017136	0.01108703	adcy5	ENSDARG00000091342
gcat	849.996923	0.57978821	0.15446462	3.75353404	0.00017436	0.01114755	gcat	ENSDARG00000005643
evpla	6995.76351	-0.3764892	0.10037106	-3.7509735	0.00017615	0.01114755	evpla	ENSDARG00000019808
ccna2	21097.0197	0.3490289	0.09301054	3.75257374	0.00017503	0.01114755	ccna2	ENSDARG00000011094
stom	1345.2423	-0.5093229	0.13606552	-3.743218	0.00018168	0.01141415	stom	ENSDARG00000003835
golgb1	3439.31789	-0.3850714	0.10316353	-3.7326314	0.00018949	0.01181714	golgb1	ENSDARG00000061951
kirrel3l	510.921867	-0.6340799	0.16995492	-3.730871	0.00019082	0.01181714	kirrel3l	ENSDARG00000104665
chek2	477.810937	-0.6057022	0.16262641	-3.7245006	0.0001957	0.01195714	chek2	ENSDARG00000025820
zgc:101744	1929.32648	0.41354992	0.11104021	3.72432586	0.00019584	0.01195714	zgc:101744	ENSDARG00000038694
oip5-as1	2418.36067	-0.4065987	0.10932528	-3.7191646	0.00019988	0.01211876	oip5-as1	ENSDARG00000093456
ENSDARG00000104184	672.321429	0.57582253	0.15558394	3.70104091	0.00021472	0.01275065	NA	ENSDARG00000104184
zgc:91849	3018.31493	-0.3749213	0.10126729	-3.7022938	0.00021366	0.01275065	zgc:91849	ENSDARG00000104716
zgc:110540	976.091105	-0.5123915	0.1383675	-3.70312	0.00021296	0.01275065	zgc:110540	ENSDARG00000054929
si:rp71-45k5.2	227.128423	0.87598828	0.236831	3.69879067	0.00021663	0.0127767	si:rp71-45k5.2	ENSDARG00000093129
oga	6634.3617	-0.3865464	0.10464671	-3.693823	0.00022091	0.01294101	oga	ENSDARG00000074686
gpat4	2061.91227	-0.4513599	0.12226732	-3.6915825	0.00022286	0.01296794	gpat4	ENSDARG00000019897
acot14	303.142676	-0.7364022	0.19972885	-3.6870097	0.0002269	0.013109	acot14	ENSDARG00000089159
hirip3	2023.15182	-0.451305	0.12255191	-3.6825618	0.0002309	0.013109	hirip3	ENSDARG00000027749
crlf3	1040.43423	0.51989131	0.14119486	3.68208393	0.00023134	0.013109	crlf3	ENSDARG00000070261
erbb2	747.12372	-0.5259621	0.14274011	-3.6847533	0.00022892	0.013109	erbb2	ENSDARG00000026294

arid4a	2694.30206	-0.3966631	0.10787347	-3.6771146	0.00023589	0.01328014	arid4a	ENSDARG00000043873
desi2	1050.78174	0.49026136	0.1334327	3.67422212	0.00023858	0.01334482	desi2	ENSDARG00000004460
kmt2bb	3484.1336	-0.4009191	0.10932357	-3.6672707	0.00024515	0.01353808	kmt2bb	ENSDARG00000060697
LOC103909544	1256.60041	0.46210215	0.12600096	3.66744942	0.00024498	0.01353808	LOC103909544	ENSDARG00000093787
yipf5	1255.64626	0.51485618	0.1405105	3.66418295	0.00024813	0.01361571	yipf5	ENSDARG00000007279
ttf2	395.455813	-0.6484888	0.1775197	-3.6530525	0.00025914	0.013955	ttf2	ENSDARG00000104105
serpinb14	281.231147	-0.7798315	0.21345756	-3.6533328	0.00025886	0.013955	serpinb14	ENSDARG00000091801
slc39a5	1498.39862	-0.4268302	0.11682673	-3.6535324	0.00025866	0.013955	slc39a5	ENSDARG00000079525
lrp6	2551.44077	-0.3868324	0.10618695	-3.6429374	0.00026954	0.01406894	lrp6	ENSDARG00000100143
crybg1b	828.4588	-0.5053209	0.13882363	-3.6400206	0.00027262	0.01406894	crybg1b	ENSDARG00000006060
si:dkey-56m19.5	6517.82376	-0.3692099	0.10141179	-3.6406997	0.0002719	0.01406894	si:dkey-56m19.5	ENSDARG00000068432
notch3	2007.65618	-0.4914745	0.13481911	-3.6454365	0.00026694	0.01406894	notch3	ENSDARG00000052139
pi4k2b	716.367546	0.69137874	0.18963491	3.64584101	0.00026652	0.01406894	pi4k2b	ENSDARG00000013881
gpx8	601.424111	0.59922719	0.16421893	3.64895325	0.00026331	0.01406894	gpx8	ENSDARG00000013302
pak4	3973.18044	0.36517308	0.10022848	3.64340648	0.00026905	0.01406894	pak4	ENSDARG00000018110
prdm14	345.516655	-0.703859	0.193695	-3.6338523	0.00027922	0.01432452	prdm14	ENSDARG00000045371
ahctf1	5235.96501	-0.3585888	0.09877732	-3.630275	0.00028312	0.01443909	ahctf1	ENSDARG00000077530
exd3	1250.95427	0.43854767	0.12099483	3.62451566	0.0002895	0.01467833	exd3	ENSDARG00000063140
otud4	1920.57541	0.43735215	0.12091528	3.6170131	0.00029802	0.0150224	otud4	ENSDARG00000077810
rad51	1060.63845	-0.4770205	0.13199473	-3.6139356	0.00030158	0.01502721	rad51	ENSDARG00000041411
ND4L	1094.0497	0.51859587	0.14349621	3.61400388	0.0003015	0.01502721	ND4L	ENSDARG00000063916
tmem176l.3a	169.433496	-1.2368929	0.34253396	-3.6110082	0.00030501	0.01511102	tmem176l.3a	ENSDARG00000098387
slkb	1361.24852	-0.4631542	0.12855059	-3.6028946	0.00031469	0.01550221	slkb	ENSDARG00000012574
si:ch211-149a19.3	591.294985	-0.5603468	0.15567584	-3.5994462	0.0003189	0.01562048	si:ch211-149a19.3	ENSDARG00000079456
msrb1a	993.892117	0.49455912	0.13753806	3.59579824	0.0003234	0.01566404	msrb1a	ENSDARG00000025436
cdc45	478.378216	-0.6465199	0.1797967	-3.5958387	0.00032335	0.01566404	cdc45	ENSDARG00000043720
her7	1068.49304	-0.5117662	0.1424954	-3.5914575	0.00032883	0.01583883	her7	ENSDARG00000017917
ENSDARG00000094597	758.164568	0.59427189	0.16586873	3.58278433	0.00033995	0.01615236	NA	ENSDARG00000094597
mboat1	1317.45956	-0.4551128	0.12705438	-3.5820319	0.00034093	0.01615236	mboat1	ENSDARG00000029356
ppp1r2	1272.40452	0.45582152	0.12720409	3.58338724	0.00033917	0.01615236	ppp1r2	ENSDARG00000054007
wdr18	1656.38537	0.43265159	0.12095003	3.5771102	0.00034741	0.0162815	wdr18	ENSDARG00000041113

sox2	730.602773	-0.6525168	0.18237392	-3.5779062	0.00034636	0.0162815	sox2	ENSDARG00000070913
ENSDARG00000008275	487.139692	-0.7556647	0.21144397	-3.5738295	0.0003518	0.01639833	NA	ENSDARG00000008275
natd1	2318.07997	-0.4184756	0.11723379	-3.5695816	0.00035755	0.0165774	natd1	ENSDARG00000038281
ENSDARG00000104551	1117.41845	0.62278587	0.17456464	3.567652	0.00036019	0.01661109	NA	ENSDARG00000104551
acacb	2353.93379	-0.3850812	0.10800523	-3.5653941	0.00036331	0.01666612	acacb	ENSDARG00000061994
ENSDARG00000089961	173.126945	-0.9045761	0.2548687	-3.5491848	0.00038643	0.01684451	NA	ENSDARG00000089961
phka1a	221.683173	0.81311662	0.22892094	3.5519538	0.00038238	0.01684451	phka1a	ENSDARG00000105159
ppp1r12c	528.345663	0.57790666	0.16283429	3.54904769	0.00038663	0.01684451	ppp1r12c	ENSDARG00000052423
leng8	1892.46428	-0.4345013	0.12220187	-3.5556027	0.00037711	0.01684451	leng8	ENSDARG00000076805
snrpd3	4900.40207	0.37041126	0.10410022	3.55821773	0.00037338	0.01684451	snrpd3	ENSDARG00000013800
piwil1	1045.25037	0.5001873	0.14062634	3.55685346	0.00037532	0.01684451	piwil1	ENSDARG00000041699
memo1	937.132353	0.46397839	0.13067213	3.55070664	0.0003842	0.01684451	memo1	ENSDARG00000010823
ric8b	661.127007	0.69499705	0.19537426	3.55726	0.00037474	0.01684451	ric8b	ENSDARG00000005972
si:ch73-299h12.2	492.013003	0.65343602	0.18353988	3.56018556	0.00037059	0.01684451	si:ch73-299h12.2	ENSDARG00000102731
si:ch211-225g23.1	1334.62754	-0.5083546	0.14306096	-3.5534127	0.00038027	0.01684451	si:ch211-225g23.1	ENSDARG00000091271
ehbp1	1063.72377	0.48718701	0.13732555	3.54767936	0.00038864	0.01684759	ehbp1	ENSDARG00000043643
incenp	4361.16172	-0.358627	0.10133748	-3.5389373	0.00040174	0.01732884	incenp	ENSDARG00000099194
prdx3	967.485135	0.4578874	0.1298277	3.52688522	0.00042048	0.01804729	prdx3	ENSDARG00000032102
ccdc106b	388.647696	0.7256343	0.20601603	3.5222225	0.00042794	0.01827725	ccdc106b	ENSDARG00000058578
cuedc2	1087.0971	0.50042341	0.14255139	3.51047731	0.0004473	0.01901038	cuedc2	ENSDARG00000039365
gcnt4a	269.928379	-0.730898	0.20834711	-3.5080784	0.00045136	0.01908906	gcnt4a	ENSDARG00000035198
asap2b	845.23285	-0.5274922	0.15042717	-3.5066282	0.00045382	0.01910021	asap2b	ENSDARG00000019564
tatdn2	1834.20323	0.46993456	0.13408112	3.50485245	0.00045686	0.01913517	tatdn2	ENSDARG00000070618
eif4ea	596.751125	-0.5537325	0.15844986	-3.4946862	0.00047462	0.01978341	eif4ea	ENSDARG00000077012
iqgap1	2855.01188	0.38539327	0.11035488	3.49230848	0.00047886	0.01986487	iqgap1	ENSDARG00000078888
nat8l	299.227907	-0.9259118	0.26598338	-3.4810889	0.00049938	0.02004389	nat8l	ENSDARG00000077256
shcbp1	3277.11705	0.49053765	0.14090687	3.48128982	0.00049901	0.02004389	shcbp1	ENSDARG00000102068
si:dkey-92f12.2	350.82714	-0.7142986	0.20504894	-3.4835519	0.00049481	0.02004389	si:dkey-92f12.2	ENSDARG00000086490
dbf4	1395.1723	-0.4098349	0.11764956	-3.4835223	0.00049486	0.02004389	dbf4	ENSDARG00000074796
snx25	603.442584	-0.5435912	0.15591661	-3.4864227	0.00048953	0.02004389	snx25	ENSDARG00000099545
nectin3b	492.760029	0.67773514	0.19475929	3.47986033	0.00050168	0.02004389	nectin3b	ENSDARG00000006604

cnot8	2703.66521	0.41055729	0.11794728	3.48085428	0.00049982	0.02004389	cnot8	ENSDARG00000020043
metrn	1720.51655	-0.4707057	0.13501238	-3.4863888	0.00048959	0.02004389	metrn	ENSDARG00000030367
dusp6	3108.64561	-0.3718666	0.10694596	-3.4771449	0.00050678	0.02015512	dusp6	ENSDARG00000070914
chd9	419.199055	-0.6992498	0.20126316	-3.4743059	0.00051218	0.02027659	chd9	ENSDARG00000074498
med25	2883.6948	0.36588978	0.10537283	3.47233519	0.00051595	0.02033319	med25	ENSDARG00000038005
camk2g1	1659.65273	0.42769631	0.12324795	3.47021031	0.00052005	0.020402	camk2g1	ENSDARG00000071395
sub1b	1481.69454	0.43065701	0.12431595	3.46421373	0.00053178	0.02076831	sub1b	ENSDARG00000007720
cdk5rap2	568.85517	0.54152861	0.15663172	3.45733675	0.00054554	0.02121011	cdk5rap2	ENSDARG00000024219
ypel3	3210.80128	0.35659321	0.10324774	3.45376304	0.00055282	0.02139721	ypel3	ENSDARG00000055510
si:dkey-13i19.8	1251.25441	0.42297165	0.12260617	3.44983996	0.00056092	0.02151844	si:dkey-13i19.8	ENSDARG00000093058
cxcr3.1	280.984845	0.75116435	0.21770105	3.45043973	0.00055967	0.02151844	cxcr3.1	ENSDARG00000007358
sugt1	1994.87271	0.41575739	0.12065994	3.44569522	0.00056959	0.02156492	sugt1	ENSDARG00000100083
mthfr	589.540151	0.56230111	0.16314324	3.44667132	0.00056754	0.02156492	mthfr	ENSDARG00000053087
topbp1	1142.48709	-0.4691391	0.13607104	-3.4477513	0.00056527	0.02156492	topbp1	ENSDARG00000059322
pank1a	276.56677	0.75067348	0.21794014	3.4444021	0.00057232	0.02156968	pank1a	ENSDARG00000008192
brca2	1758.51143	-0.3937969	0.11436663	-3.4432847	0.00057469	0.02156968	brca2	ENSDARG00000079015
ENSDARG00000026166	319.028156	0.6646859	0.19325436	3.43943554	0.00058293	0.02173267	NA	ENSDARG00000026166
rbmx2	1138.52313	-0.528261	0.15361268	-3.4389153	0.00058405	0.02173267	rbmx2	ENSDARG00000044380
greb1l	656.400186	-0.5686651	0.1654201	-3.4377027	0.00058667	0.02173693	greb1l	ENSDARG00000039196
ifi30	1845.1144	-0.4231613	0.12315643	-3.4359661	0.00059044	0.02178365	ifi30	ENSDARG00000056378
ccdc187	718.40044	-0.4941418	0.14456979	-3.4180158	0.00063079	0.0230759	ccdc187	ENSDARG00000053857
cldn7b	4535.63529	0.39304199	0.11498469	3.4182116	0.00063034	0.0230759	cldn7b	ENSDARG00000014047
pkd2	710.078746	-0.506152	0.14826735	-3.4137791	0.00064068	0.02333923	pkd2	ENSDARG00000014098
ptdss1a	1570.94722	0.39120778	0.11477553	3.40845977	0.00065331	0.02364258	ptdss1a	ENSDARG00000012588
abcb10	622.090901	-0.5114626	0.15007808	-3.4079763	0.00065447	0.02364258	abcb10	ENSDARG00000061591
ywhaqb	3674.93071	-0.3664979	0.10762402	-3.4053545	0.00066078	0.02373186	ywhaqb	ENSDARG00000023323
l2hgdh	761.967757	0.47506649	0.1395333	3.40468183	0.00066241	0.02373186	l2hgdh	ENSDARG00000060500
selenot1b	970.679713	0.46661076	0.1371215	3.40290003	0.00066675	0.02378886	selenot1b	ENSDARG00000027595
rsrp1	2415.06708	-0.4046978	0.11915486	-3.3964014	0.00068278	0.02426113	rsrp1	ENSDARG00000030440
glceb	620.604544	0.59991507	0.17699405	3.38946469	0.00070029	0.02478177	glceb	ENSDARG00000068981
ENSDARG00000103751	222.159857	-0.7586019	0.22455815	-3.3781981	0.00072962	0.02571483	NA	ENSDARG00000103751

rgs3b	797.751192	-0.490442	0.1454529	-3.3718269	0.00074671	0.02610486	rgs3b	ENSDARG00000035132
myo9b	959.666141	-0.4736815	0.14046652	-3.372202	0.0007457	0.02610486	myo9b	ENSDARG00000077410
rbm5	3861.05534	-0.3324237	0.09866275	-3.369293	0.00075361	0.02624025	rbm5	ENSDARG00000098280
nxn	415.749577	0.59262508	0.17595834	3.36798526	0.0007572	0.02625957	nxn	ENSDARG00000033978
cited2	660.52125	0.52198934	0.15513217	3.36480389	0.00076598	0.02645838	cited2	ENSDARG00000030905
eve1	1105.93612	-0.6043729	0.17968912	-3.3634362	0.00076979	0.02648431	eve1	ENSDARG00000056012
si:dkey-102m7.3	933.114536	-0.4747754	0.14126298	-3.3609329	0.0007768	0.02661988	si:dkey-102m7.3	ENSDARG00000100366
slc35c1	942.679195	-0.485737	0.14487532	-3.3527929	0.00080001	0.02707027	slc35c1	ENSDARG00000104669
fthl27	6779.90335	-0.3258178	0.09712231	-3.3547165	0.00079446	0.02707027	fthl27	ENSDARG00000031776
gmds	2222.2425	-0.3915956	0.116826	-3.3519555	0.00080243	0.02707027	gmds	ENSDARG00000026629
tmem258	520.08072	0.54903799	0.16377622	3.35236702	0.00080124	0.02707027	tmem258	ENSDARG00000078785
ewsr1a	8180.17703	-0.3481403	0.10395763	-3.3488668	0.00081143	0.02726776	ewsr1a	ENSDARG00000020258
tnip1	1362.71483	-0.4023253	0.12024429	-3.3458989	0.00082016	0.02734928	tnip1	ENSDARG00000015653
baiap2l1a	1730.94758	0.41934586	0.12531304	3.34638649	0.00081872	0.02734928	baiap2l1a	ENSDARG00000029305
rpl14	4157.86323	-0.3666526	0.10964442	-3.344015	0.00082575	0.02743018	rpl14	ENSDARG00000103433
ergic3	1926.33811	-0.3725059	0.1116055	-3.3377021	0.00084474	0.02795389	ergic3	ENSDARG00000038074
pxna	1212.5197	0.4388405	0.13164024	3.33363498	0.00085719	0.02825796	pxna	ENSDARG00000088590
si:dkey-238c7.16	1249.99707	0.43487658	0.13059853	3.32987348	0.00086885	0.02853397	si:dkey-238c7.16	ENSDARG00000069537
lamtor4	1178.52932	0.41762808	0.12551697	3.327264	0.00087703	0.02853942	lamtor4	ENSDARG00000045542
tsc1a	1268.33445	0.46506753	0.13979961	3.32667265	0.0008789	0.02853942	tsc1a	ENSDARG00000026048
afp4	4441.91826	-0.3507661	0.10543937	-3.3267093	0.00087878	0.02853942	afp4	ENSDARG00000095863
gsnb	615.022054	0.5032635	0.15147748	3.32236511	0.00089258	0.02887557	gsnb	ENSDARG00000045262
itpka	162.016203	-0.8408507	0.25384133	-3.3125053	0.00092464	0.02969134	itpka	ENSDARG00000042856
lmo2	297.452537	-0.8208858	0.2477897	-3.3128325	0.00092356	0.02969134	lmo2	ENSDARG00000095019
rras	2389.74101	0.37310262	0.11281895	3.30709187	0.0009427	0.03000078	rras	ENSDARG00000006553
fmnl2b	2978.68867	-0.4210212	0.127331	-3.3065096	0.00094466	0.03000078	fmnl2b	ENSDARG00000075041
naa50	1527.00842	0.40790767	0.12328915	3.30854472	0.00093782	0.03000078	naa50	ENSDARG00000027825
LOC110437815	967.141139	-0.5087332	0.1540947	-3.3014321	0.00096193	0.03028927	LOC110437815	ENSDARG00000103441
atp6v0a2b	445.458123	0.61685204	0.1868817	3.30076218	0.00096423	0.03028927	atp6v0a2b	ENSDARG00000035565
mgat4a	540.775355	0.5685527	0.17219708	3.30175568	0.00096082	0.03028927	mgat4a	ENSDARG00000063330
dbr1	2651.59208	0.34279998	0.10400073	3.29613055	0.00098026	0.03067512	dbr1	ENSDARG00000056923

rab3gap2	2242.89956	0.36778737	0.11161373	3.29518042	0.00098359	0.03067512	rab3gap2	ENSDARG00000044136
ubr5	9033.19598	-0.3876623	0.11792117	-3.2874703	0.00101092	0.03141457	ubr5	ENSDARG00000018192
hltf	727.396177	-0.4927991	0.15005984	-3.284017	0.00102339	0.0316508	hltf	ENSDARG00000026053
fibpb	261.660647	-0.8047997	0.24511562	-3.2833473	0.00102582	0.0316508	fibpb	ENSDARG00000020811
smap1	1582.68388	0.39877393	0.12150455	3.2819669	0.00103086	0.03169337	smap1	ENSDARG00000031302
znf185	7832.80265	-0.3036286	0.09270227	-3.2753097	0.00105546	0.0322674	znf185	ENSDARG00000103917
maco1b	978.974032	0.4950924	0.15117759	3.27490599	0.00105697	0.0322674	maco1b	ENSDARG00000012741
ENSDARG00000035770	2358.14116	-0.3787463	0.11580777	-3.2704739	0.00107367	0.03227842	NA	ENSDARG00000035770
atg5	224.86466	0.83550254	0.25591059	3.26482207	0.00109533	0.03227842	atg5	ENSDARG00000023396
glipr2l	534.040601	0.50787076	0.15556628	3.26465836	0.00109596	0.03227842	glipr2l	ENSDARG00000016837
mark3a	904.282475	0.47111841	0.14425102	3.26596237	0.00109093	0.03227842	mark3a	ENSDARG00000019345
dna2	924.186782	-0.4342324	0.13280816	-3.2696212	0.00107692	0.03227842	dna2	ENSDARG00000078759
ppp1r37	928.649234	0.51099918	0.15633486	3.26861965	0.00108073	0.03227842	ppp1r37	ENSDARG00000078458
nrip1a	630.773101	-0.4954607	0.1516275	-3.2676179	0.00108457	0.03227842	nrip1a	ENSDARG00000068965
vps13d	910.834671	0.43602586	0.13319596	3.2735667	0.00106199	0.03227842	vps13d	ENSDARG00000017986
si:ch211-69b7.6	1084.60866	-0.4494973	0.13771116	-3.2640584	0.00109829	0.03227842	si:ch211-69b7.6	ENSDARG00000102146
eps15l1a	508.602113	-0.5441804	0.16669565	-3.2645148	0.00109652	0.03227842	eps15l1a	ENSDARG00000042670
ND1	6863.41989	0.40731476	0.12456304	3.26994888	0.00107567	0.03227842	ND1	ENSDARG00000063895
ppp1cc	3001.80083	0.40151622	0.12310649	3.26153581	0.0011081	0.03245697	ppp1cc	ENSDARG00000099226
alcamb	8088.79657	-0.3095282	0.09503587	-3.2569618	0.00112612	0.03287349	alcamb	ENSDARG00000058538
gapdh	2520.4896	0.35581781	0.10931221	3.25505993	0.00113368	0.03298337	gapdh	ENSDARG00000043457
pds5b	2076.30359	-0.3875136	0.11922651	-3.2502301	0.00115312	0.03299512	pds5b	ENSDARG00000098897
gigyf1b	3335.472	-0.3400966	0.10463621	-3.2502761	0.00115293	0.03299512	gigyf1b	ENSDARG00000078691

Appendix 3 – *sox32* mutant 9.00 hpf top 300 DEGs

Gene Name	baseMean	log2FoldChange	lfcSE	stat	pvalue	padj	symbol	Emsebl_name
txn	658.903732	-3.8494808	0.20303156	-18.960012	3.65E-80	7.04E-76	txn	ENSDARG00000044125
sox17	353.332519	-7.0470481	0.40274077	-17.497727	1.49E-68	1.44E-64	sox17	ENSDARG00000101717
ENSDARG00000093936	135.457304	-6.3553669	0.52692368	-12.061266	1.69E-33	1.09E-29	NA	ENSDARG00000093936
lrrc17	106.376592	3.40464969	0.32642997	10.4299543	1.81E-25	8.72E-22	lrrc17	ENSDARG00000037960
abcb11a	1294.12354	2.45766194	0.25246533	9.73465132	2.15E-22	8.27E-19	abcb11a	ENSDARG00000011573
sbf2	225.284672	2.49091157	0.26454768	9.41573763	4.70E-21	1.51E-17	sbf2	ENSDARG00000059460
wfs1b	298.76493	-2.0058468	0.21492493	-9.3327785	1.03E-20	2.84E-17	wfs1b	ENSDARG00000074617
gnsb	2163.54864	1.0247786	0.11432293	8.96389415	3.13E-19	7.55E-16	gnsb	ENSDARG00000098296
arl3l2	285.717431	-2.3342503	0.26506272	-8.8064076	1.29E-18	2.77E-15	arl3l2	ENSDARG00000015404
slc30a1b	934.698039	-1.226393	0.14257374	-8.6018156	7.85E-18	1.37E-14	slc30a1b	ENSDARG00000053896
psme1	84.5569775	-3.0622687	0.35573131	-8.6083754	7.41E-18	1.37E-14	psme1	ENSDARG00000002165
lrrc15	679.258129	1.60060072	0.18875747	8.47966806	2.26E-17	3.63E-14	lrrc15	ENSDARG00000070792
lad1	317.225478	-2.3171734	0.27995358	-8.2769917	1.26E-16	1.87E-13	lad1	ENSDARG00000022698
prex1	422.414385	-1.6979778	0.20811713	-8.1587602	3.38E-16	4.66E-13	prex1	ENSDARG00000075793
slc38a7	1423.06151	-1.4010094	0.17346371	-8.0766715	6.66E-16	8.55E-13	slc38a7	ENSDARG00000012002
s1pr5a	763.241525	-1.2184186	0.1517435	-8.0294613	9.79E-16	1.18E-12	s1pr5a	ENSDARG00000040526
pltp	388.642142	-1.5404328	0.19340379	-7.964853	1.65E-15	1.88E-12	pltp	ENSDARG00000104495
hsd3b2	4269.55761	-1.3089284	0.16548239	-7.9097751	2.58E-15	2.76E-12	hsd3b2	ENSDARG00000019747
slmapb	122.352035	-2.3990372	0.30430274	-7.8837186	3.18E-15	3.22E-12	slmapb	ENSDARG00000020764
si:dkey-69o16.5	480.146904	-1.3290768	0.16885244	-7.8712326	3.51E-15	3.38E-12	si:dkey-69o16.5	ENSDARG00000070057
ENSDARG00000104440	49.68862	-6.1903595	0.79652236	-7.7717335	7.74E-15	7.11E-12	NA	ENSDARG00000104440
slco1d1	667.660858	1.08164144	0.14162886	7.63715421	2.22E-14	1.95E-11	slco1d1	ENSDARG00000104108
anxa13	186.271552	-1.974889	0.25912015	-7.6215185	2.51E-14	2.10E-11	anxa13	ENSDARG00000013976
anpepb	50.0964685	-3.4613115	0.45987958	-7.5265607	5.21E-14	4.18E-11	anpepb	ENSDARG00000103878
zgc:112994	5937.33107	1.07957366	0.14383196	7.50579839	6.11E-14	4.71E-11	zgc:112994	ENSDARG00000026296
atg16l2	72.9338705	-2.5040633	0.3339648	-7.4979856	6.48E-14	4.80E-11	atg16l2	ENSDARG00000043238
zgc:92380	79.9420641	-2.4607021	0.32983839	-7.4603265	8.63E-14	6.16E-11	zgc:92380	ENSDARG00000041339
trim2b	105.509757	-2.3653045	0.31776978	-7.4434533	9.81E-14	6.75E-11	trim2b	ENSDARG00000076174

si:ch211-146m13.3	45.2395371	-4.0392124	0.54523972	-7.4081404	1.28E-13	8.51E-11	si:ch211-146m13.3	ENSDARG00000059534
cnnm2b	170.862641	-1.563895	0.21961441	-7.1210946	1.07E-12	6.88E-10	cnnm2b	ENSDARG00000078733
trim16	872.341849	-1.265455	0.17831016	-7.0969314	1.28E-12	7.93E-10	trim16	ENSDARG00000010673
tbc1d24	1121.42001	0.97157491	0.13858836	7.0105088	2.37E-12	1.43E-09	tbc1d24	ENSDARG00000069339
nwd2	95.1156604	2.28987335	0.32782513	6.98504527	2.85E-12	1.66E-09	nwd2	ENSDARG00000077162
si:ch73-74h11.1	261.758564	2.22762263	0.32304102	6.8957888	5.36E-12	3.04E-09	si:ch73-74h11.1	ENSDARG00000062750
si:dkey-28e7.3	158.552171	-1.6829442	0.24422882	-6.8908503	5.55E-12	3.05E-09	si:dkey-28e7.3	ENSDARG00000074508
dgat2	290.817603	-1.5763691	0.23058297	-6.8364505	8.12E-12	4.35E-09	dgat2	ENSDARG00000018846
ENSDARG00000104302	31.723043	4.73330772	0.69397385	6.8205851	9.07E-12	4.72E-09	NA	ENSDARG00000104302
sorbs2b	70.7958487	-2.5652538	0.37806562	-6.7852077	1.16E-11	5.88E-09	sorbs2b	ENSDARG00000061603
mid1ip1a	1955.5902	-1.6169513	0.23845108	-6.7810612	1.19E-11	5.90E-09	mid1ip1a	ENSDARG00000041051
prrt4	40.4495072	-4.0899915	0.60791343	-6.7279177	1.72E-11	8.29E-09	prrt4	ENSDARG00000088343
apoc2	1348.90115	1.24233262	0.18618458	6.67258586	2.51E-11	1.18E-08	apoc2	ENSDARG00000092155
lrrc4ba	52.1336728	-8.1683909	1.23308313	-6.6243635	3.49E-11	1.60E-08	lrrc4ba	ENSDARG00000004597
csf3a	48.9173004	-8.0751242	1.23776327	-6.5239649	6.85E-11	3.07E-08	csf3a	ENSDARG00000102211
hcn5	120.144894	-1.8922599	0.29194032	-6.4816671	9.07E-11	3.89E-08	hcn5	ENSDARG00000077382
smpd2a	1199.6159	-1.0051113	0.15503385	-6.4831728	8.98E-11	3.89E-08	smpd2a	ENSDARG00000040523
si:dkey-33c14.3	27.3380335	-8.2037859	1.27284051	-6.4452584	1.15E-10	4.84E-08	si:dkey-33c14.3	ENSDARG00000097118
cdh6	3275.49348	-0.9272696	0.14396417	-6.4409747	1.19E-10	4.85E-08	cdh6	ENSDARG00000014522
tspan2b	136.528068	2.15914006	0.33535061	6.43845569	1.21E-10	4.85E-08	tspan2b	ENSDARG00000059202
slc26a4	29.7833604	-3.9527778	0.6170348	-6.4060857	1.49E-10	5.87E-08	slc26a4	ENSDARG00000069431
foxa2	1001.47897	-0.8511922	0.13352642	-6.3747097	1.83E-10	7.07E-08	foxa2	ENSDARG00000003411
zgc:86709	52.3822585	3.3345584	0.52419284	6.36132002	2.00E-10	7.56E-08	zgc:86709	ENSDARG00000057911
ankrd6a	293.084646	-1.5593016	0.24751719	-6.2997711	2.98E-10	1.10E-07	ankrd6a	ENSDARG00000057790
fbxl7	54.7621366	-2.5207827	0.40422435	-6.2360982	4.49E-10	1.63E-07	fbxl7	ENSDARG00000062251
cfi	1556.37119	-0.7885432	0.12655228	-6.2309679	4.64E-10	1.65E-07	cfi	ENSDARG00000099425
gpm6ab	110.876	-1.7240789	0.27760419	-6.2105653	5.28E-10	1.85E-07	gpm6ab	ENSDARG00000004621
bicc2	26.0218988	-4.6590677	0.76358144	-6.1015989	1.05E-09	3.61E-07	bicc2	ENSDARG00000076557
mixl1	130.870442	1.42579619	0.23464295	6.07645024	1.23E-09	4.15E-07	mixl1	ENSDARG00000069252
prdx5	982.071573	-0.8566737	0.14252157	-6.0108352	1.85E-09	6.13E-07	prdx5	ENSDARG00000055064
dpp7	3769.44242	-0.9413759	0.15736175	-5.9822411	2.20E-09	7.19E-07	dpp7	ENSDARG00000027750

LOC100329294	68.1051793	-2.2292777	0.37283334	-5.9792872	2.24E-09	7.20E-07	LOC100329294	ENSDARG00000098478
dhcr7	80.9496615	-2.2995542	0.38528407	-5.9684644	2.39E-09	7.57E-07	dhcr7	ENSDARG00000103226
fdps	571.770185	-0.9307271	0.15651947	-5.9463986	2.74E-09	8.52E-07	fdps	ENSDARG00000040890
ENSDARG00000093964	70.754854	-2.0500057	0.34541856	-5.9348453	2.94E-09	9.00E-07	NA	ENSDARG00000093964
ENSDARG00000023609	127.520681	1.45991539	0.24613577	5.9313419	3.00E-09	9.05E-07	NA	ENSDARG00000023609
pvalb9	310.033003	1.61251588	0.27249937	5.91750317	3.27E-09	9.69E-07	pvalb9	ENSDARG00000071601
mfsd4ab	25.2131488	-4.6854563	0.79376477	-5.9028273	3.57E-09	1.04E-06	mfsd4ab	ENSDARG00000008263
si:ch73-234b20.5	141.486609	-1.6549378	0.28152013	-5.8785771	4.14E-09	1.19E-06	si:ch73-234b20.5	ENSDARG00000086996
nudt4b	1373.14503	-0.8745929	0.14924556	-5.8600932	4.63E-09	1.31E-06	nudt4b	ENSDARG00000045878
mnx2b	23.4093248	-4.2884385	0.7349429	-5.8350635	5.38E-09	1.50E-06	mnx2b	ENSDARG00000030350
slc23a3	102.271217	-2.6025502	0.44703531	-5.8218001	5.82E-09	1.60E-06	slc23a3	ENSDARG00000088891
sesn3	7550.97316	0.88085495	0.15348753	5.7389348	9.53E-09	2.56E-06	sesn3	ENSDARG00000015822
scarb1	1026.82108	0.76362403	0.13307957	5.73810092	9.57E-09	2.56E-06	scarb1	ENSDARG00000101557
syne3	996.70243	0.87322509	0.15447442	5.65287819	1.58E-08	4.17E-06	syne3	ENSDARG00000023237
acp5a	2178.93966	0.67733606	0.12078162	5.60794005	2.05E-08	5.33E-06	acp5a	ENSDARG00000019763
mc5rb	22.9022554	-4.3947067	0.78577114	-5.5928583	2.23E-08	5.74E-06	mc5rb	ENSDARG00000054946
prkacba	285.267801	-1.0643875	0.19186681	-5.5475329	2.90E-08	7.35E-06	prkacba	ENSDARG00000001782
zgc:114041	292.321929	-1.4108739	0.25585405	-5.51437	3.50E-08	8.76E-06	zgc:114041	ENSDARG00000052109
pla1a	176.959773	-1.4531699	0.26530658	-5.4773235	4.32E-08	1.07E-05	pla1a	ENSDARG00000102176
got1	594.589009	0.92014015	0.16833871	5.46600454	4.60E-08	1.12E-05	got1	ENSDARG00000039093
si:ch1073-390k14.1	139.161914	-1.6371536	0.30132577	-5.4331683	5.54E-08	1.33E-05	si:ch1073-390k14.1	ENSDARG00000088549
sall3a	121.133748	1.35178774	0.25120421	5.38123046	7.40E-08	1.76E-05	sall3a	ENSDARG00000079613
fam162a	151.455895	-1.2932938	0.2417106	-5.3505881	8.77E-08	2.06E-05	fam162a	ENSDARG00000063344
mttp	11443.0728	0.66257814	0.12396612	5.34483237	9.05E-08	2.10E-05	mttp	ENSDARG00000008637
nlrc3	282.427091	1.00273135	0.18797933	5.33426383	9.59E-08	2.18E-05	nlrc3	ENSDARG00000103146
mafb	261.277475	-1.0639842	0.19947692	-5.3338712	9.61E-08	2.18E-05	mafb	ENSDARG00000076520
ENSDARG00000100894	278.127779	1.3256476	0.25056998	5.29052843	1.22E-07	2.73E-05	NA	ENSDARG00000100894
ENSDARG00000088187	39.9430377	-2.6594996	0.50307725	-5.2864637	1.25E-07	2.76E-05	NA	ENSDARG00000088187
scarb2a	2149.04418	0.78541267	0.14974591	5.24496893	1.56E-07	3.42E-05	scarb2a	ENSDARG00000098312
si:ch211-11p18.6	43.0753953	2.17106442	0.4149688	5.23187385	1.68E-07	3.63E-05	si:ch211-11p18.6	ENSDARG00000077068
phospho1	2513.98117	-0.6636917	0.12719857	-5.2177605	1.81E-07	3.88E-05	phospho1	ENSDARG00000008403

g6pca.2	18.8823549	-4.4267987	0.84908757	-5.213595	1.85E-07	3.92E-05	g6pca.2	ENSDARG00000013721
stmn4l	128.283547	-1.3092558	0.25266135	-5.1818602	2.20E-07	4.60E-05	stmn4l	ENSDARG00000043932
LOC100334085	467.992368	1.09958017	0.21252937	5.17377995	2.29E-07	4.75E-05	LOC100334085	ENSDARG00000099747
snai1b	1101.01153	0.65244823	0.12734227	5.12357942	3.00E-07	6.15E-05	snai1b	ENSDARG00000046019
tarsl2	240.381853	1.01594528	0.19860776	5.11533522	3.13E-07	6.31E-05	tarsl2	ENSDARG00000092774
csgalnact1a	110.672315	-2.2844436	0.44665391	-5.1145721	3.14E-07	6.31E-05	csgalnact1a	ENSDARG00000040535
cacnb2b	78.3038306	1.73635605	0.34058425	5.09816901	3.43E-07	6.81E-05	cacnb2b	ENSDARG00000055565
vtg7	12.6464466	-7.0930043	1.39546454	-5.0828982	3.72E-07	7.31E-05	vtg7	ENSDARG00000092419
acap3a	438.781333	-1.1682807	0.23075639	-5.0628316	4.13E-07	8.04E-05	acap3a	ENSDARG00000075990
anks4b	159.688473	1.10352206	0.21806267	5.06057303	4.18E-07	8.06E-05	anks4b	ENSDARG00000036846
LOC110437952	11.4093165	-6.9404002	1.37788682	-5.0369886	4.73E-07	9.02E-05	LOC110437952	ENSDARG00000097762
cxcr4a	1491.66905	-0.8350592	0.16607094	-5.0283286	4.95E-07	9.35E-05	cxcr4a	ENSDARG00000057633
prkag2a	108.916743	-1.3374278	0.2670508	-5.0081401	5.50E-07	0.00010284	prkag2a	ENSDARG00000012625
noctb	93.6168278	1.34554719	0.27008619	4.98191772	6.30E-07	0.00011668	noctb	ENSDARG00000078525
epha3	29.4111356	-2.5178741	0.50562538	-4.9797225	6.37E-07	0.00011688	epha3	ENSDARG00000039373
ENSDARG00000098531	54.5835638	1.91281192	0.38564086	4.96008621	7.05E-07	0.00012812	NA	ENSDARG00000098531
jag1a	171.545179	-1.1228854	0.22666824	-4.9538718	7.28E-07	0.00013105	jag1a	ENSDARG00000030289
eps8l3a	176.614426	1.16880497	0.23607377	4.95101573	7.38E-07	0.00013175	eps8l3a	ENSDARG00000101979
dok2	16.8622739	4.30389676	0.87518533	4.91769755	8.76E-07	0.00015484	dok2	ENSDARG00000075818
cthl	81.7686269	-1.5733792	0.32064483	-4.9069222	9.25E-07	0.00016211	cthl	ENSDARG00000032206
zgc:101744	636.327201	-0.7573778	0.15519015	-4.880321	1.06E-06	0.00018391	zgc:101744	ENSDARG00000038694
kifc3	693.234029	-0.7141098	0.14680237	-4.8644298	1.15E-06	0.00019579	kifc3	ENSDARG00000054978
aqp12	166.386864	-1.5061292	0.30953471	-4.8657844	1.14E-06	0.00019579	aqp12	ENSDARG00000043279
angptl7	132.872264	-2.1097653	0.43649105	-4.8334674	1.34E-06	0.00022685	angptl7	ENSDARG00000027582
nt5c2l1	643.080141	0.91417959	0.18943876	4.8257263	1.39E-06	0.00023379	nt5c2l1	ENSDARG00000034852
fabp11a	29.9466673	2.6608194	0.55263741	4.81476525	1.47E-06	0.00024487	fabp11a	ENSDARG00000017299
qpct	197.67998	1.28153163	0.26660295	4.80689224	1.53E-06	0.00025253	qpct	ENSDARG00000089717
exoc3l1	220.818325	-0.9186237	0.19142469	-4.7988777	1.60E-06	0.00026062	exoc3l1	ENSDARG00000051899
c8b	10.3885843	-6.8100592	1.4219314	-4.7893022	1.67E-06	0.00027107	c8b	ENSDARG00000039517
prpsap1	495.946591	-0.8032688	0.16817125	-4.7764933	1.78E-06	0.00028651	prpsap1	ENSDARG00000099222
fam160b2	271.937612	0.91436966	0.19194498	4.76370709	1.90E-06	0.00030276	fam160b2	ENSDARG00000060029

rbp4l	84.7133938	-1.3921435	0.2934413	-4.7441975	2.09E-06	0.00033071	rbp4l	ENSDARG00000044684
hsd17b12a	3202.46812	-0.6262464	0.13216784	-4.7382664	2.16E-06	0.00033777	hsd17b12a	ENSDARG00000015709
iqca1	56.5554954	-1.6728485	0.35326237	-4.7354278	2.19E-06	0.00033977	iqca1	ENSDARG00000057276
si:ch211-191i18.4	467.112981	0.96809542	0.20564927	4.70750721	2.51E-06	0.00038666	si:ch211-191i18.4	ENSDARG00000095328
elovl7b	1496.64266	-0.9097726	0.19383107	-4.6936368	2.68E-06	0.00041055	elovl7b	ENSDARG00000100185
rrm2b	1635.08525	0.61635337	0.1319845	4.66989203	3.01E-06	0.00045735	rrm2b	ENSDARG00000033367
serpinh1b	3823.09669	-0.7532333	0.16151066	-4.6636754	3.11E-06	0.00046771	serpinh1b	ENSDARG00000019949
abtb2a	557.000602	-1.0263377	0.22119278	-4.6400143	3.48E-06	0.00052053	abtb2a	ENSDARG00000059751
rerg	193.23392	-1.4029896	0.30321193	-4.6270924	3.71E-06	0.00054981	rerg	ENSDARG00000104632
cyp2aa3	1593.97103	-1.048242	0.22748121	-4.6080376	4.06E-06	0.00059806	cyp2aa3	ENSDARG00000103347
irx7	3192.33265	-0.5651747	0.12306207	-4.5925989	4.38E-06	0.0006392	irx7	ENSDARG00000002601
ENSDARG00000013390	27.9785915	-3.0299982	0.66189355	-4.5777726	4.70E-06	0.00068104	NA	ENSDARG00000013390
ntrk2a	402.850298	1.31702433	0.28899194	4.55730464	5.18E-06	0.00074527	ntrk2a	ENSDARG00000059897
si:dkey-222f8.3	1108.32643	-0.5896	0.13038393	-4.5220294	6.12E-06	0.00087446	si:dkey-222f8.3	ENSDARG00000058366
met	90.4285521	-1.6511058	0.36788788	-4.4880679	7.19E-06	0.00101857	met	ENSDARG00000070903
ugt5g1	105.043964	-1.4149697	0.31554162	-4.4842569	7.32E-06	0.00102938	ugt5g1	ENSDARG00000032862
ENSDARG00000020812	44.2521513	1.94539889	0.43528621	4.46924074	7.85E-06	0.00109635	NA	ENSDARG00000020812
pde6a	69.6294667	-1.9404788	0.43506746	-4.4601791	8.19E-06	0.00113552	pde6a	ENSDARG00000000380
zgc:174153	785.203193	0.76229613	0.17196121	4.4329539	9.30E-06	0.00127967	zgc:174153	ENSDARG00000079376
ca2	73.1547694	-1.5132172	0.34256992	-4.4172506	1.00E-05	0.00136646	ca2	ENSDARG00000014488
ENSDARG00000108628	35.2570272	2.29706325	0.52025899	4.41523032	1.01E-05	0.00136957	NA	ENSDARG00000108628
slc4a5	49.103914	2.13179566	0.48354413	4.40868896	1.04E-05	0.00140172	slc4a5	ENSDARG00000104387
shc1	1030.53377	-0.5592948	0.12747047	-4.3876418	1.15E-05	0.00153371	shc1	ENSDARG00000075437
eps8a	264.422953	-0.8675262	0.1980372	-4.3806227	1.18E-05	0.00155163	eps8a	ENSDARG00000102128
si:ch211-241j12.3	548.209969	-0.6826923	0.15578738	-4.3822055	1.17E-05	0.00155163	si:ch211-241j12.3	ENSDARG00000043963
foxi3a	160.081117	-0.9892559	0.22581483	-4.380828	1.18E-05	0.00155163	foxi3a	ENSDARG00000055926
slc22a16	115.283718	1.30772285	0.30030822	4.35460224	1.33E-05	0.00173608	slc22a16	ENSDARG00000015869
cftr	12.6957394	-5.1833708	1.19526634	-4.3365823	1.45E-05	0.00187197	cftr	ENSDARG00000041107
bco1	145.108392	-1.0916099	0.25195464	-4.3325654	1.47E-05	0.00188169	bco1	ENSDARG00000104256
sec14l1	110.027656	-1.2662744	0.29227273	-4.3325096	1.47E-05	0.00188169	sec14l1	ENSDARG00000103991
si:ch211-12p12.2	45.5537146	1.96516192	0.45477269	4.321196	1.55E-05	0.0019678	si:ch211-12p12.2	ENSDARG00000102219

adsl	1429.30234	-0.6529666	0.15121677	-4.3180835	1.57E-05	0.00196983	adsl	ENSDARG00000017049
nr1d2b	110.657325	1.10842214	0.25666556	4.31854651	1.57E-05	0.00196983	nr1d2b	ENSDARG00000009594
dtwd2	365.230887	-0.7042577	0.16335692	-4.3111594	1.62E-05	0.00201943	dtwd2	ENSDARG00000101873
diabloa	1340.46723	-0.8042188	0.18849342	-4.2665619	1.99E-05	0.0024526	diabloa	ENSDARG00000104172
gprc5c	109.070473	-1.4455589	0.33970928	-4.2552824	2.09E-05	0.00256314	gprc5c	ENSDARG00000100862
gdpd1	652.233558	-0.6812191	0.16037355	-4.2477026	2.16E-05	0.00262104	gdpd1	ENSDARG00000017261
ENSDARG00000077719	310.127653	-0.8905878	0.2096761	-4.2474457	2.16E-05	0.00262104	NA	ENSDARG00000077719
kidins220a	220.404248	-1.3183141	0.31052941	-4.2453759	2.18E-05	0.00262882	kidins220a	ENSDARG00000031240
pacs1a	88.0639201	-1.4695316	0.34646522	-4.2414981	2.22E-05	0.00265805	pacs1a	ENSDARG00000044556
aprt	418.432408	0.74484705	0.17630654	4.22472737	2.39E-05	0.00284625	aprt	ENSDARG00000003519
c3a.1	151.556166	-1.5067088	0.35679896	-4.222851	2.41E-05	0.00285245	c3a.1	ENSDARG00000012694
elovl6l	637.354785	-0.9342486	0.22156117	-4.2166622	2.48E-05	0.00291396	elovl6l	ENSDARG00000038639
ftr86	762.179255	-0.8984979	0.21399467	-4.1986927	2.68E-05	0.00313594	ftr86	ENSDARG00000076839
tex36	14.6476974	-3.5695586	0.85136732	-4.1927362	2.76E-05	0.00320006	tex36	ENSDARG00000097920
cald1b	28.1052894	2.2338546	0.53364785	4.18600878	2.84E-05	0.00327661	cald1b	ENSDARG00000086391
trip10a	231.377705	-0.8419059	0.20174435	-4.1731323	3.00E-05	0.00340629	trip10a	ENSDARG00000005679
ENSDARG00000035367	6.85170718	-6.2081762	1.48749637	-4.173574	3.00E-05	0.00340629	NA	ENSDARG00000035367
apbb1ip	192.635067	-0.8566261	0.20519402	-4.1747127	2.98E-05	0.00340629	apbb1ip	ENSDARG00000016505
fabp2	2813.1182	-0.9812873	0.23574116	-4.1625625	3.15E-05	0.00354705	fabp2	ENSDARG00000006427
mansc1	153.123577	-1.3791129	0.33167099	-4.1580752	3.21E-05	0.0035964	mansc1	ENSDARG00000104839
trim3a	158.936034	1.26446017	0.30493904	4.14659977	3.37E-05	0.00375953	trim3a	ENSDARG00000063711
il12rb2l	1018.02918	-0.6972105	0.16826552	-4.1435141	3.42E-05	0.0037886	il12rb2l	ENSDARG00000074850
mmel1	80.7507356	1.852826	0.45077561	4.11030672	3.95E-05	0.00435189	mmel1	ENSDARG00000105389
olfm1b	14.0762638	3.52718637	0.86061805	4.09843409	4.16E-05	0.00455518	olfm1b	ENSDARG00000014053
cant1b	1157.96642	-0.5750423	0.14068314	-4.0874996	4.36E-05	0.00474824	cant1b	ENSDARG00000102977
ampd2b	905.888976	-0.5780853	0.14217352	-4.0660544	4.78E-05	0.00517754	ampd2b	ENSDARG00000029952
zgc:194678	11.4009285	-4.9908349	1.22969751	-4.0585874	4.94E-05	0.00531601	zgc:194678	ENSDARG00000098810
golim4a	396.467966	-0.7404648	0.1825172	-4.0569591	4.97E-05	0.00532345	golim4a	ENSDARG00000100977
ypel3	885.345246	0.62199595	0.15342567	4.05405398	5.03E-05	0.00536026	ypel3	ENSDARG00000055510
LOC110439320	130.365205	-1.0837033	0.26760706	-4.0496064	5.13E-05	0.00543313	LOC110439320	ENSDARG00000098821
ENSDARG00000104777	113.432322	-1.0493829	0.25982751	-4.0387676	5.37E-05	0.00565926	NA	ENSDARG00000104777

adgrf3b	57.2907813	-1.4491012	0.35902503	-4.0362122	5.43E-05	0.00569013	adgrf3b	ENSDARG00000093008
zgc:162171	52.129329	-1.4023526	0.3479706	-4.0300893	5.58E-05	0.00580884	zgc:162171	ENSDARG00000036499
si:ch211-256m1.8	12.741226	3.60263371	0.89591299	4.02118705	5.79E-05	0.00600039	si:ch211-256m1.8	ENSDARG00000055172
si:ch211-156j16.1	88.7502664	-1.2508027	0.31168702	-4.0130087	5.99E-05	0.00617899	si:ch211-156j16.1	ENSDARG00000092035
zgc:101562	127.194003	-1.3187988	0.32993997	-3.9970871	6.41E-05	0.00657435	zgc:101562	ENSDARG00000040179
rgs3a	172.604182	0.84349932	0.21137438	3.9905466	6.59E-05	0.00672257	rgs3a	ENSDARG00000099746
glula	10371.1652	-0.4906473	0.12304544	-3.987529	6.68E-05	0.00677279	glula	ENSDARG00000099776
atp8a2	543.42922	-1.8927408	0.4752866	-3.9823148	6.82E-05	0.0068869	atp8a2	ENSDARG00000077492
slc12a5b	30.9171605	-1.8920275	0.47544443	-3.9794924	6.91E-05	0.00689695	slc12a5b	ENSDARG00000078187
znf710a	814.391147	-0.5799289	0.14568823	-3.9806159	6.87E-05	0.00689695	znf710a	ENSDARG00000014680
slc6a1b	5.86143431	-5.979509	1.51509518	-3.9466227	7.93E-05	0.00787465	slc6a1b	ENSDARG00000039647
pnp4a	536.637913	0.92981505	0.2358849	3.94181672	8.09E-05	0.00799295	pnp4a	ENSDARG00000057575
cryba2b	12.2125192	-4.1164004	1.04615502	-3.9347901	8.33E-05	0.00818845	cryba2b	ENSDARG00000041925
LOC562097	5.87164028	-5.9818465	1.52450612	-3.9237931	8.72E-05	0.008485	LOC562097	ENSDARG00000059529
ENSDARG00000089107	166.957794	1.08361436	0.27609829	3.92474132	8.68E-05	0.008485	NA	ENSDARG00000089107
igfn1.1	10.3022108	-4.8668013	1.24177739	-3.9192221	8.88E-05	0.00860407	igfn1.1	ENSDARG00000005526
sp8b	757.004431	-0.5473427	0.13979823	-3.9152335	9.03E-05	0.00870383	sp8b	ENSDARG00000056666
cbfa2t3	13.8945091	-3.078877	0.78763254	-3.9090272	9.27E-05	0.00888604	cbfa2t3	ENSDARG00000079012
rasl11b	1139.16543	-0.5274427	0.13540499	-3.8952975	9.81E-05	0.00935822	rasl11b	ENSDARG00000015611
zgc:110699	365.69475	-0.6805593	0.17531944	-3.8818247	0.00010368	0.00984356	zgc:110699	ENSDARG00000017474
dennd2db	37.3898782	1.66893994	0.43187893	3.86436993	0.00011138	0.01052288	dennd2db	ENSDARG00000030250
os9	181.897543	-1.1586203	0.30070813	-3.8529728	0.00011669	0.01097136	os9	ENSDARG00000020301
si:dkey-205h13.2	10.3045242	-4.2275957	1.10324053	-3.8319801	0.00012712	0.01189337	si:dkey-205h13.2	ENSDARG00000089429
pck2	14070.7695	-0.6500843	0.16984954	-3.8274125	0.0001295	0.01205765	pck2	ENSDARG00000020956
rrm2	2476.16453	-0.8912919	0.23318644	-3.8222289	0.00013225	0.01225482	rrm2	ENSDARG00000078069
selenop2	889.743814	0.97516108	0.25580964	3.81205751	0.00013781	0.01270929	selenop2	ENSDARG00000079727
osgn1	125.60603	-1.1708956	0.30851637	-3.7952462	0.0001475	0.01353741	osgn1	ENSDARG00000052279
hagh	214.63566	1.14133114	0.30098406	3.79199858	0.00014944	0.01361321	hagh	ENSDARG00000025338
cxcl12b	2042.03256	0.51141181	0.13488349	3.79150787	0.00014974	0.01361321	cxcl12b	ENSDARG00000055100
si:dkey-92f12.2	58.7910005	-1.241913	0.328633	-3.7790272	0.00015744	0.01424667	si:dkey-92f12.2	ENSDARG00000086490
pou6f2	25.1449504	1.94864568	0.51662875	3.77184907	0.00016204	0.0145944	pou6f2	ENSDARG00000086362

plin2	167.099692	0.90233443	0.23933778	3.7701295	0.00016316	0.01462698	plin2	ENSDARG00000042332
pde11al	264.386644	-3.3521497	0.89275024	-3.7548572	0.00017344	0.01547634	pde11al	ENSDARG00000006151
atrn	757.860853	0.5667833	0.15162233	3.73812561	0.0001854	0.01646704	atrn	ENSDARG00000062164
atp1b3a	2050.61223	0.42930417	0.11505973	3.73114198	0.00019061	0.0167397	atp1b3a	ENSDARG00000015790
rictorb	606.425084	-0.6656802	0.17833714	-3.7327062	0.00018943	0.0167397	rictorb	ENSDARG00000002020
eya1	144.387288	0.97974729	0.26262907	3.73053626	0.00019107	0.0167397	eya1	ENSDARG00000014259
ggps1	750.0073	-0.6475516	0.17447376	-3.7114556	0.00020607	0.01797199	ggps1	ENSDARG00000023627
aanat2	33.3342379	1.85200231	0.49916649	3.71018956	0.0002071	0.01798074	aanat2	ENSDARG00000079802
dlb	126.549931	0.96212316	0.25950948	3.70746829	0.00020934	0.01805049	dlb	ENSDARG00000004232
xirp1	41.2210885	-1.8871988	0.50909914	-3.7069377	0.00020978	0.01805049	xirp1	ENSDARG00000030722
agpat2	734.434049	0.81961201	0.22134791	3.70282251	0.00021321	0.01826439	agpat2	ENSDARG00000101139
apoa1b	21170.2334	-0.7231467	0.19538852	-3.7010702	0.00021469	0.01830962	apoa1b	ENSDARG00000101324
si:dkeyp-27e10.3	155.039121	1.19480063	0.3230693	3.69827967	0.00021707	0.01842557	si:dkeyp-27e10.3	ENSDARG00000063008
rras2	436.43149	0.58714019	0.15883624	3.69651271	0.00021858	0.01842557	rras2	ENSDARG00000036252
ENSDARG00000106519	8.72907249	5.61626727	1.51950351	3.69611998	0.00021892	0.01842557	NA	ENSDARG00000106519
ednraa	450.742446	-0.71803	0.19469115	-3.6880464	0.00022598	0.01893733	ednraa	ENSDARG00000011876
gxylt1b	238.399631	0.86150285	0.23386535	3.68375578	0.00022982	0.01917575	gxylt1b	ENSDARG00000022550
ppp1r14bb	1432.446	0.51654917	0.14071058	3.67100437	0.0002416	0.01999138	ppp1r14bb	ENSDARG00000030161
slc44a5b	181.93369	0.79473947	0.21649557	3.67092715	0.00024167	0.01999138	slc44a5b	ENSDARG00000057419
foxj1a	39.3842142	-1.5382211	0.42030125	-3.6598062	0.00025241	0.02079006	foxj1a	ENSDARG00000101919
flnb	44.5123632	-1.9427111	0.53135622	-3.6561369	0.00025604	0.02100002	flnb	ENSDARG00000092281
st3gal3b	259.002729	-0.6744673	0.18488671	-3.6480028	0.00026429	0.02158417	st3gal3b	ENSDARG00000015252
ENSDARG00000088836	29.1577563	-1.9060279	0.52279879	-3.6458153	0.00026655	0.02167677	NA	ENSDARG00000088836
mthfd1l	810.914519	0.703461	0.193069	3.64357303	0.00026888	0.02177471	mthfd1l	ENSDARG00000042221
ahi1	591.242808	-0.6711063	0.18454527	-3.6365404	0.00027632	0.02228399	ahi1	ENSDARG00000044056
si:ch211-130m23.2	249.564889	0.90654944	0.24959171	3.63212967	0.00028109	0.02257401	si:ch211-130m23.2	ENSDARG00000095136
agrn	45.9525643	1.6962795	0.46727493	3.63015302	0.00028325	0.02265321	agrn	ENSDARG00000079388
rwdd	552.395747	-0.7288687	0.20185855	-3.6107893	0.00030527	0.02422914	rwdd	ENSDARG00000068256
nfil3-5	830.857546	0.568346	0.15740978	3.61061419	0.00030547	0.02422914	nfil3-5	ENSDARG00000094965
tfeb	177.502401	0.79026456	0.21901813	3.60821528	0.00030831	0.02435401	tfeb	ENSDARG00000010794
si:dkey-175m17.7	205.128406	-0.7713922	0.21398639	-3.6048658	0.00031231	0.02456957	si:dkey-175m17.7	ENSDARG00000078317

sfmbt1	265.082756	0.71899307	0.19967979	3.60073038	0.00031732	0.02486225	sfmbt1	ENSDARG00000044915
gcshb	221.77892	-0.6780098	0.18848986	-3.5970624	0.00032183	0.02511327	gcshb	ENSDARG00000105187
tm2d3	422.710916	-0.6049495	0.16853162	-3.5895313	0.00033127	0.02574579	tm2d3	ENSDARG00000076618
ifi30	3600.99353	0.54036892	0.15086907	3.58170774	0.00034136	0.02642284	ifi30	ENSDARG00000056378
zgc:101040	157.810531	0.86005312	0.24034225	3.57845165	0.00034564	0.02664714	zgc:101040	ENSDARG00000005176
cpne7	8.50209941	4.68083183	1.30891793	3.57610796	0.00034875	0.02677993	cpne7	ENSDARG00000102584
ENSDARG00000107898	65.4413666	-1.3211687	0.36973585	-3.5732773	0.00035254	0.0269638	NA	ENSDARG00000107898
cdh12a	5.09795908	-5.7815793	1.62207841	-3.5643032	0.00036482	0.02779296	cdh12a	ENSDARG00000078226
rasa1b	153.859234	0.89388049	0.25088002	3.56298004	0.00036667	0.0278235	rasa1b	ENSDARG00000073665
arhgap36	423.948563	-0.684316	0.19223937	-3.5597078	0.00037127	0.02795239	arhgap36	ENSDARG00000059672
aoc2	1593.99509	-0.6704548	0.18832256	-3.5601409	0.00037066	0.02795239	aoc2	ENSDARG00000014646
pcbp4	184.065357	-1.380184	0.38852775	-3.5523434	0.00038182	0.02863473	pcbp4	ENSDARG00000024276
ENSDARG00000096607	9.2773426	4.16316221	1.17275895	3.5498874	0.0003854	0.02868001	NA	ENSDARG00000096607
chrna2a	141.574848	-3.8958068	1.0974178	-3.549976	0.00038527	0.02868001	chrna2a	ENSDARG00000006602
si:dkey-157l19.2	773.400957	0.52111911	0.14730218	3.53775558	0.00040354	0.02991498	si:dkey-157l19.2	ENSDARG00000060340
si:ch1073-155h21.1	210.807693	-1.3117751	0.37105525	-3.5352555	0.00040738	0.03008376	si:ch1073-155h21.1	ENSDARG00000007582
actn3b	449.077944	-0.72137	0.20439725	-3.5292548	0.00041673	0.03042726	actn3b	ENSDARG00000001431
tgm2l	42.9372792	-1.7278868	0.48968558	-3.5285638	0.00041782	0.03042726	tgm2l	ENSDARG00000093381
krt99	12531.7528	-0.3851932	0.10916539	-3.5285288	0.00041788	0.03042726	krt99	ENSDARG00000019365
cry2	312.130176	0.67863614	0.1923446	3.52823081	0.00041835	0.03042726	cry2	ENSDARG00000102403
cfb	52.8836284	-1.8311328	0.51918144	-3.5269613	0.00042036	0.03045859	cfb	ENSDARG00000055278
xkr5b	827.979372	-0.5782932	0.16449165	-3.5156388	0.0004387	0.03166836	xkr5b	ENSDARG00000097530
asb11	2001.28931	-0.4624842	0.13201945	-3.5031516	0.00045979	0.03294405	asb11	ENSDARG00000056561
tspan13a	16.4856783	2.22439597	0.63496955	3.50315375	0.00045978	0.03294405	tspan13a	ENSDARG00000068883
slc13a4	2529.85818	0.46680134	0.13351939	3.4961314	0.00047206	0.03369782	slc13a4	ENSDARG00000059053
wipi1	118.276417	-1.0313321	0.29520981	-3.4935562	0.00047663	0.03389895	wipi1	ENSDARG00000040657
znf185	4175.28846	-0.5813196	0.16686918	-3.4836847	0.00049456	0.03504479	znf185	ENSDARG00000103917
zfpm2b	14.8854001	2.73147969	0.78449353	3.4818384	0.00049798	0.03515804	zfpm2b	ENSDARG00000100560
tysnd1	35.8476762	1.52115839	0.43731052	3.47843997	0.00050434	0.03547692	tysnd1	ENSDARG00000074895
adpgk2	176.254445	0.86001441	0.24777457	3.47095514	0.00051861	0.03621631	adpgk2	ENSDARG00000062785
lamp2	3265.8755	0.44371553	0.12782644	3.47123425	0.00051807	0.03621631	lamp2	ENSDARG00000014914

ENSDARG00000098082	104.094119	1.21169659	0.34919491	3.46997213	0.00052051	0.03621789	NA	ENSDARG00000098082
bckdhbl	1446.34365	0.44275748	0.12795224	3.46033396	0.00053951	0.03740444	bckdhbl	ENSDARG00000093569
uncx4.1	35.9449363	-1.4463031	0.41846344	-3.4562234	0.0005478	0.03784343	uncx4.1	ENSDARG00000037760
sgsm3	101.111498	-0.9055837	0.26222277	-3.4534901	0.00055338	0.03795693	sgsm3	ENSDARG00000019038
LOC100007813	10.2578876	-3.2811154	0.94982825	-3.4544302	0.00055146	0.03795693	LOC100007813	ENSDARG00000087999
prph2b	20.5401412	2.46260645	0.71650572	3.43696689	0.00058827	0.04020662	prph2b	ENSDARG00000014840
ENSDARG00000087345	42.8311887	-1.3755113	0.40286839	-3.4142944	0.00063947	0.04355204	NA	ENSDARG00000087345
rybpb	585.109472	0.56650027	0.16614573	3.40965889	0.00065044	0.04406896	rybpb	ENSDARG00000053459
slc29a1a	2202.94079	0.5178954	0.15191299	3.40915807	0.00065164	0.04406896	slc29a1a	ENSDARG00000101289
ftr14l	489.622237	-0.5324376	0.15628037	-3.4069386	0.00065696	0.04427353	ftr14l	ENSDARG00000078254
klf3	512.798317	-0.5920343	0.17409863	-3.4005687	0.00067246	0.04497488	klf3	ENSDARG00000015495
osbpl7	1037.01376	-0.5593969	0.16452953	-3.3999785	0.00067391	0.04497488	osbpl7	ENSDARG00000012981
fam126b	311.887737	-0.7314632	0.21514928	-3.3997939	0.00067437	0.04497488	fam126b	ENSDARG00000100400
cd36	484.551064	0.74265405	0.21955769	3.38250069	0.00071829	0.04773162	cd36	ENSDARG00000032639
msto1	1788.9843	0.47084137	0.13923634	3.38159817	0.00072065	0.04773162	msto1	ENSDARG00000024381
si:dkey-229b18.3	403.380805	-0.5918766	0.17511535	-3.3799239	0.00072506	0.04785887	si:dkey-229b18.3	ENSDARG00000095912
grm8a	39.4625532	1.57736289	0.4673349	3.37523021	0.00073754	0.04835149	grm8a	ENSDARG00000077654
cyfip2	206.564873	0.97005189	0.28734045	3.37596698	0.00073557	0.04835149	cyfip2	ENSDARG00000036375
dazap2	520.612269	0.51971573	0.15440942	3.36582914	0.00076314	0.04986018	dazap2	ENSDARG00000007867
cyp2k16	140.127613	-1.1295131	0.33579508	-3.3636976	0.00076906	0.0499344	cyp2k16	ENSDARG00000102981
hhla2b.1	15.8216129	-2.5507959	0.75836324	-3.3635542	0.00076946	0.0499344	hhla2b.1	ENSDARG00000088680

Appendix 4 – *sox17:gfp* top 300 DEGs

Gene Name	baseMean	log2FoldChange	lfcSE	stat	pvalue	padj	symbol	Emsebl_name
abtb2a	223.487841	3.55199728	0.21331713	16.6512523	2.96E-62	5.74E-58	abtb2a	ENSDARG00000059751
si:ch211-146m13.3	228.740921	5.16244264	0.35127763	14.6961897	6.82E-49	6.61E-45	si:ch211-146m13.3	ENSDARG00000059534
vwf	238.812969	6.85871527	0.47110879	14.5586653	5.15E-48	3.32E-44	vwf	ENSDARG00000077231
iqca1	201.426148	3.90434852	0.29881288	13.0661987	5.14E-39	2.49E-35	iqca1	ENSDARG00000057276
sox17	2925.53981	6.13200821	0.48077957	12.7543028	2.95E-37	1.14E-33	sox17	ENSDARG00000101717
cdh6	1722.09362	2.89717688	0.23968925	12.0872205	1.23E-33	3.99E-30	cdh6	ENSDARG00000014522
si:ch73-234b20.5	118.930732	3.94077956	0.32685143	12.0567915	1.79E-33	4.95E-30	si:ch73-234b20.5	ENSDARG00000086996
ssuh2rs1	258.313089	2.38507461	0.20063906	11.8873895	1.38E-32	3.33E-29	ssuh2rs1	ENSDARG00000063292
slc26a4	112.584492	4.84905897	0.41824222	11.5939013	4.43E-31	9.53E-28	slc26a4	ENSDARG00000069431
pltp	1171.72135	2.94373229	0.25642884	11.479724	1.67E-30	3.23E-27	pltp	ENSDARG00000104495
gpm6ab	513.91232	3.94617946	0.34945531	11.2923724	1.43E-29	2.52E-26	gpm6ab	ENSDARG00000004621
gata5	1687.36343	1.96707268	0.18181468	10.8191082	2.79E-27	4.51E-24	gata5	ENSDARG00000017821
mnx2b	98.7060724	5.79669743	0.54993149	10.5407629	5.60E-26	8.36E-23	mnx2b	ENSDARG00000030350
met	356.868131	3.33965637	0.3219815	10.3721997	3.32E-25	4.59E-22	met	ENSDARG00000070903
sox32	1209.16338	4.32099036	0.43420271	9.9515509	2.48E-23	3.21E-20	sox32	ENSDARG00000100591
sept9a	853.083371	1.72964541	0.17602957	9.82588001	8.71E-23	1.06E-19	sept9a	ENSDARG00000020235
prex1	661.335177	2.34548272	0.25000621	9.38169788	6.49E-21	7.40E-18	prex1	ENSDARG00000075793
tfa	221.050848	2.61268215	0.27937307	9.35194688	8.60E-21	9.27E-18	tfa	ENSDARG00000016771
scarb2b	96.0179468	3.3870143	0.36579586	9.25930186	2.06E-20	2.10E-17	scarb2b	ENSDARG00000100753
prrt4	46.4921834	3.8727977	0.42232953	9.17008494	4.73E-20	4.58E-17	prrt4	ENSDARG00000088343
lhfpl6	283.665559	2.26042408	0.25776193	8.76942578	1.80E-18	1.66E-15	lhfpl6	ENSDARG00000004363
flrt3	2740.89559	1.55604217	0.17868063	8.70851049	3.08E-18	2.71E-15	flrt3	ENSDARG00000076895
prdx5	646.955078	1.87800751	0.21710764	8.65012159	5.14E-18	4.34E-15	prdx5	ENSDARG00000055064
zgc:193505	131.049053	-2.3319294	0.27024041	-8.6290923	6.18E-18	4.99E-15	zgc:193505	ENSDARG00000093584
col9a2	45.8606332	3.6906549	0.44289608	8.33300419	7.88E-17	6.11E-14	col9a2	ENSDARG00000024492
prkacba	36.4477959	4.13545882	0.50016255	8.26822969	1.36E-16	1.01E-13	prkacba	ENSDARG00000001782
si:dkey-95h12.1	87.633034	3.51433985	0.43815175	8.0208281	1.05E-15	7.54E-13	si:dkey-95h12.1	ENSDARG00000040100
ENSDARG00000102395	77.4347516	2.68164008	0.33527346	7.99836673	1.26E-15	8.73E-13	NA	ENSDARG00000102395

csnk2a2a	412.351133	0.93241875	0.11685866	7.97902982	1.47E-15	9.86E-13	csnk2a2a	ENSDARG00000012818
ENSDARG00000089986	219.019998	1.55595939	0.19526567	7.96842279	1.61E-15	1.04E-12	NA	ENSDARG00000089986
cdh12a	39.8374047	6.29149565	0.79289952	7.93479563	2.11E-15	1.32E-12	cdh12a	ENSDARG00000078226
mcama	628.153235	1.89976271	0.2400709	7.91334036	2.51E-15	1.52E-12	mcama	ENSDARG00000089643
krt5	2024.33954	-1.8597499	0.23520417	-7.9069598	2.64E-15	1.55E-12	krt5	ENSDARG00000058371
foxj1a	90.7090348	2.32086309	0.295399	7.85670591	3.94E-15	2.20E-12	foxj1a	ENSDARG00000101919
hapln1b	143.702905	2.28756317	0.29118444	7.85606257	3.96E-15	2.20E-12	hapln1b	ENSDARG00000068516
celf3a	153.476941	1.1482035	0.15022851	7.64304648	2.12E-14	1.14E-11	celf3a	ENSDARG00000034668
zgc:194678	173.438732	5.36133284	0.70223046	7.63471988	2.26E-14	1.19E-11	zgc:194678	ENSDARG00000098810
krt17	83.3036158	-2.3360398	0.30986979	-7.5387787	4.74E-14	2.42E-11	krt17	ENSDARG00000094041
pde4bb	29.2634789	5.46935256	0.73833092	7.40772522	1.28E-13	6.27E-11	pde4bb	ENSDARG00000074233
dnah5	76.3572003	3.16851705	0.42778345	7.4068248	1.29E-13	6.27E-11	dnah5	ENSDARG00000087373
LOC100331300	30.3050437	7.14252644	0.96810836	7.37781712	1.61E-13	7.61E-11	LOC100331300	ENSDARG00000098915
arhgap36	635.033258	1.75894565	0.24017814	7.32350427	2.42E-13	1.11E-10	arhgap36	ENSDARG00000059672
ptprga	45.0808135	1.95944098	0.26842082	7.29988439	2.88E-13	1.30E-10	ptprga	ENSDARG00000045006
bicc2	48.34704	4.5531243	0.6253725	7.28065963	3.32E-13	1.46E-10	bicc2	ENSDARG00000076557
ENSDARG00000106359	234.455265	1.06683141	0.14664363	7.27499318	3.46E-13	1.49E-10	NA	ENSDARG00000106359
eml2	88.669401	2.91173443	0.40128836	7.25596531	3.99E-13	1.68E-10	eml2	ENSDARG00000008808
si:ch211-156j16.1	399.224229	1.6031564	0.2211244	7.2500205	4.17E-13	1.72E-10	si:ch211-156j16.1	ENSDARG00000092035
trpm4a	249.777334	2.63845047	0.36424104	7.24369354	4.37E-13	1.76E-10	trpm4a	ENSDARG00000059993
ponzr5	40.3204741	-2.3491767	0.32569971	-7.2127074	5.49E-13	2.17E-10	ponzr5	ENSDARG00000055046
ugp2a	33.3935457	2.71825737	0.37767775	7.19729293	6.14E-13	2.38E-10	ugp2a	ENSDARG00000005578
nav2b	1711.74034	1.04707172	0.14667003	7.13896167	9.40E-13	3.57E-10	nav2b	ENSDARG00000001879
st6galnac	1306.09976	1.37321577	0.19259023	7.13024613	1.00E-12	3.73E-10	st6galnac	ENSDARG00000086292
rxrgb	108.95308	2.56911727	0.36123189	7.11209974	1.14E-12	4.18E-10	rxrgb	ENSDARG00000004697
znf185	179.997172	-1.6556492	0.23315149	-7.1011738	1.24E-12	4.36E-10	znf185	ENSDARG00000103917
mctp1b	58.3861965	3.08191723	0.43386527	7.10339697	1.22E-12	4.36E-10	mctp1b	ENSDARG00000060871
nrp2a	98.4480985	2.32980157	0.33004874	7.05896207	1.68E-12	5.81E-10	nrp2a	ENSDARG00000096546
LOC563933	72.6781571	2.67071751	0.38031741	7.02233828	2.18E-12	7.42E-10	LOC563933	ENSDARG00000055786
ENSDARG00000059693	106.436044	2.2339329	0.3218808	6.94024908	3.91E-12	1.30E-09	NA	ENSDARG00000059693
kcnq2a	29.3047843	3.45083971	0.4973202	6.93886893	3.95E-12	1.30E-09	kcnq2a	ENSDARG00000075307

ENSDARG00000100981	51.2574324	3.58937223	0.52442842	6.84435111	7.68E-12	2.44E-09	NA	ENSDARG00000100981
krt99	350.987653	-1.7687753	0.25840794	-6.8448954	7.65E-12	2.44E-09	krt99	ENSDARG00000019365
tacc1	155.957528	2.16016311	0.31700287	6.81433301	9.47E-12	2.96E-09	tacc1	ENSDARG00000073753
trip10a	180.838727	1.56055764	0.22921402	6.80829913	9.88E-12	3.04E-09	trip10a	ENSDARG00000005679
flrt2	53.1872494	2.9109362	0.42815187	6.7988404	1.05E-11	3.19E-09	flrt2	ENSDARG00000079355
tbx1	387.142339	1.20141603	0.17722833	6.77891645	1.21E-11	3.61E-09	tbx1	ENSDARG00000031891
alox5b.3	208.373391	-1.5348406	0.22859794	-6.7141489	1.89E-11	5.56E-09	alox5b.3	ENSDARG00000069966
ppl	392.028585	-2.0415272	0.30506545	-6.6920957	2.20E-11	6.36E-09	ppl	ENSDARG00000101043
zgc:162707	42.179651	2.30164563	0.34422584	6.68644053	2.29E-11	6.52E-09	zgc:162707	ENSDARG00000061664
fstl1b	1403.17947	0.932326	0.13981409	6.66832665	2.59E-11	7.27E-09	fstl1b	ENSDARG00000039576
krt4	3058.52365	-1.5671956	0.23544386	-6.6563452	2.81E-11	7.77E-09	krt4	ENSDARG00000017624
st5	105.789026	1.48680465	0.22402703	6.63671981	3.21E-11	8.76E-09	st5	ENSDARG00000037363
krt97	244.046285	-1.8821341	0.28409952	-6.6249111	3.47E-11	9.35E-09	krt97	ENSDARG00000000212
LOC100334443	4187.71467	0.93047094	0.14077316	6.60971827	3.85E-11	1.02E-08	LOC100334443	ENSDARG00000040503
gpnmb	37.878979	4.91019768	0.74476554	6.59294425	4.31E-11	1.13E-08	gpnmb	ENSDARG00000062688
cyt1	69.2529815	-2.5587439	0.38828084	-6.5899308	4.40E-11	1.14E-08	cyt1	ENSDARG00000092947
ENSDARG00000057669	17.8098433	3.48516181	0.52934488	6.58391518	4.58E-11	1.17E-08	NA	ENSDARG00000057669
uox	32.5302864	-2.1001177	0.32083191	-6.5458506	5.92E-11	1.49E-08	uox	ENSDARG00000007024
abcc9	106.000805	1.98619391	0.30695427	6.47065078	9.76E-11	2.42E-08	abcc9	ENSDARG00000015985
aldh3b1	130.90523	-1.297235	0.20181622	-6.4278036	1.29E-10	3.18E-08	aldh3b1	ENSDARG00000013839
anxa1c	83.6801531	-2.0718959	0.32475392	-6.379895	1.77E-10	4.29E-08	anxa1c	ENSDARG00000104359
cftr	28.0119699	5.56474362	0.87916286	6.32959359	2.46E-10	5.88E-08	cftr	ENSDARG00000041107
nrp1b	63.9484542	2.2011532	0.34896411	6.3076779	2.83E-10	6.70E-08	nrp1b	ENSDARG00000027290
kif1aa	34.7966909	2.49461476	0.39658203	6.29028697	3.17E-10	7.40E-08	kif1aa	ENSDARG00000061817
spag6	18.7971146	5.17834486	0.82665319	6.26422895	3.75E-10	8.65E-08	spag6	ENSDARG00000020158
castor2	789.135367	0.85656355	0.13772049	6.21957955	4.98E-10	1.14E-07	castor2	ENSDARG00000018985
hhip	536.279897	1.33221224	0.21650377	6.15329811	7.59E-10	1.71E-07	hhip	ENSDARG00000060397
cyp2aa4	436.099668	1.506236	0.24538961	6.13814079	8.35E-10	1.86E-07	cyp2aa4	ENSDARG00000098803
ahcyl2	1262.11307	1.04821412	0.17214404	6.08916866	1.13E-09	2.50E-07	ahcyl2	ENSDARG00000039343
krt18a.1	1847.63973	-1.1825239	0.1944007	-6.08292	1.18E-09	2.57E-07	krt18a.1	ENSDARG00000018404
jag1a	213.011889	1.78778144	0.29522106	6.05573818	1.40E-09	3.01E-07	jag1a	ENSDARG00000030289

pkp1b	52.3682175	-1.9633696	0.32496434	-6.0418001	1.52E-09	3.25E-07	pkp1b	ENSDARG00000052705
si:ch211-157c3.4	69.1624725	-2.2138566	0.36732155	-6.0270261	1.67E-09	3.52E-07	si:ch211-157c3.4	ENSDARG00000087093
cpda	1402.76242	0.76347248	0.1272588	5.99936898	1.98E-09	4.13E-07	cpda	ENSDARG00000055648
cd9a	64.7786474	1.97055173	0.32878409	5.99345222	2.05E-09	4.24E-07	cd9a	ENSDARG00000005842
tie1	71.4247851	2.34591206	0.39430564	5.9494763	2.69E-09	5.49E-07	tie1	ENSDARG00000004105
cox4i2	70.6017353	1.9313665	0.32514546	5.94000758	2.85E-09	5.75E-07	cox4i2	ENSDARG00000022509
rbms2b	487.047069	1.07557321	0.18113512	5.93796047	2.89E-09	5.77E-07	rbms2b	ENSDARG00000056150
ak9	16.2816844	4.71008519	0.80080172	5.88171215	4.06E-09	8.03E-07	ak9	ENSDARG00000021913
daw1	20.6720467	3.8218415	0.65462798	5.83818845	5.28E-09	1.02E-06	daw1	ENSDARG00000021462
krt92	729.919487	-1.4065976	0.24090071	-5.8389103	5.25E-09	1.02E-06	krt92	ENSDARG00000036834
ENSDARG00000106792	13.764192	4.67257413	0.80447522	5.8082263	6.31E-09	1.21E-06	NA	ENSDARG00000106792
ets2	224.280988	1.66425817	0.289883	5.74113753	9.40E-09	1.79E-06	ets2	ENSDARG00000103980
hephl1a	361.403035	1.29385958	0.2257601	5.731126	9.98E-09	1.88E-06	hephl1a	ENSDARG00000059231
ENSDARG00000097207	15.6425001	4.65921374	0.81592494	5.71034601	1.13E-08	2.10E-06	NA	ENSDARG00000097207
cyt1l	133.608088	-1.5026512	0.26333643	-5.7062032	1.16E-08	2.13E-06	cyt1l	ENSDARG00000036832
mgll	81.5017771	-1.6709684	0.29356308	-5.6920253	1.26E-08	2.30E-06	mgll	ENSDARG00000036820
aoc2	2094.175	1.09524443	0.19276905	5.6816404	1.33E-08	2.42E-06	aoc2	ENSDARG00000014646
si:ch211-125o16.4	150.063641	-1.8372533	0.32438512	-5.6638026	1.48E-08	2.66E-06	si:ch211-125o16.4	ENSDARG00000056836
krt8	2844.19958	-0.916139	0.16186833	-5.6597795	1.52E-08	2.70E-06	krt8	ENSDARG00000058358
sash1a	1857.65466	0.67599723	0.11953584	5.65518445	1.56E-08	2.74E-06	sash1a	ENSDARG00000007179
cygb1	29.132179	3.1743237	0.5615905	5.65238137	1.58E-08	2.76E-06	cygb1	ENSDARG00000099371
asph	7430.629	0.73876203	0.13155543	5.61559531	1.96E-08	3.39E-06	asph	ENSDARG00000055945
wfs1b	129.812802	2.00728977	0.35871916	5.59571382	2.20E-08	3.77E-06	wfs1b	ENSDARG00000074617
slc43a2b	265.152086	1.70491007	0.30574182	5.5763064	2.46E-08	4.18E-06	slc43a2b	ENSDARG00000061120
fmnl3	391.520531	1.25139815	0.22458537	5.57203768	2.52E-08	4.24E-06	fmnl3	ENSDARG00000004372
grhl1	21.6587078	-2.3455825	0.42190893	-5.5594523	2.71E-08	4.52E-06	grhl1	ENSDARG00000061391
gstt1b	128.987034	1.55834073	0.28225211	5.52109507	3.37E-08	5.58E-06	gstt1b	ENSDARG00000017388
fam84b	22.1678686	2.99650838	0.5435852	5.5124907	3.54E-08	5.81E-06	fam84b	ENSDARG00000032859
esrrgb	18.9737291	3.02956449	0.55072265	5.50107119	3.77E-08	6.15E-06	esrrgb	ENSDARG00000011696
cd81a	3166.05854	0.7029028	0.12813427	5.48567364	4.12E-08	6.65E-06	cd81a	ENSDARG00000036080
dhrs3b	1640.06492	0.98342844	0.18004724	5.4620577	4.71E-08	7.54E-06	dhrs3b	ENSDARG00000044803

ENSDARG00000078072	26.4339124	-2.1503849	0.39510445	-5.4425732	5.25E-08	8.34E-06	NA	ENSDARG00000078072
fgfrl1b	103.219708	2.36025749	0.43404699	5.43779261	5.39E-08	8.50E-06	fgfrl1b	ENSDARG00000052556
nqo1	86.3163768	-1.4535846	0.26753334	-5.4332837	5.53E-08	8.65E-06	nqo1	ENSDARG00000010250
zgc:100920	97.2134118	1.21622787	0.22466725	5.41346319	6.18E-08	9.52E-06	zgc:100920	ENSDARG00000042961
grhl2b	460.104674	1.17247449	0.21659357	5.41324701	6.19E-08	9.52E-06	grhl2b	ENSDARG00000061974
ENSDARG00000089342	181.43462	1.00161572	0.18761974	5.33854115	9.37E-08	1.43E-05	NA	ENSDARG00000089342
anks1b	26.537233	2.26862875	0.42537096	5.33329482	9.64E-08	1.46E-05	anks1b	ENSDARG00000003512
evplb	197.608764	-2.0975456	0.39463397	-5.3151673	1.07E-07	1.60E-05	evplb	ENSDARG00000103459
tmsb1	45.9929059	-1.5827846	0.29792707	-5.3126577	1.08E-07	1.61E-05	tmsb1	ENSDARG00000104181
plekhg5b	284.813873	1.44288946	0.2717265	5.31008	1.10E-07	1.62E-05	plekhg5b	ENSDARG00000101752
anxa1b	78.3409671	-1.732157	0.32678897	-5.3005368	1.15E-07	1.70E-05	anxa1b	ENSDARG00000100095
ptgis	6.00717641	5.97200885	1.12898034	5.28973681	1.22E-07	1.79E-05	ptgis	ENSDARG00000060094
capn9	280.284089	-0.9671519	0.18308062	-5.2826559	1.27E-07	1.84E-05	capn9	ENSDARG00000012341
hsd3b2	746.467414	1.01991861	0.19517228	5.22573505	1.73E-07	2.49E-05	hsd3b2	ENSDARG00000019747
cxadr	293.476184	0.98805863	0.19025631	5.1933028	2.07E-07	2.94E-05	cxadr	ENSDARG00000043658
efhc2	12.0875472	4.52594909	0.87303916	5.18413066	2.17E-07	3.07E-05	efhc2	ENSDARG00000004204
sesn1	216.36205	1.21901029	0.23552745	5.17566112	2.27E-07	3.19E-05	sesn1	ENSDARG00000020693
mao	332.08165	1.62976675	0.31503548	5.17328002	2.30E-07	3.21E-05	mao	ENSDARG00000023712
emilin1a	9.56747778	5.46330905	1.05903749	5.15874942	2.49E-07	3.44E-05	emilin1a	ENSDARG00000024537
ENSDARG00000098058	29.4753406	-2.4153098	0.47249199	-5.1118535	3.19E-07	4.39E-05	NA	ENSDARG00000098058
tpm4a	250.404821	0.86061326	0.1697799	5.06899385	4.00E-07	5.46E-05	tpm4a	ENSDARG00000023963
plpp3	812.013572	0.87732698	0.17369537	5.050952	4.40E-07	5.96E-05	plpp3	ENSDARG00000059933
itga5	3051.28198	0.66342026	0.13138418	5.04946833	4.43E-07	5.96E-05	itga5	ENSDARG00000006353
tspan18a	30.4453243	1.69773945	0.33647915	5.04560077	4.52E-07	6.04E-05	tspan18a	ENSDARG00000056656
ENSDARG00000108014	8.76899709	6.51772051	1.29695227	5.02541277	5.02E-07	6.67E-05	NA	ENSDARG00000108014
klf11a	201.473154	1.25954705	0.2511903	5.01431399	5.32E-07	6.97E-05	klf11a	ENSDARG00000030844
gabbr1b	13.9424448	3.19295411	0.63660344	5.01560927	5.29E-07	6.97E-05	gabbr1b	ENSDARG00000016667
zmynd10	14.0080381	5.55053063	1.1081129	5.00899376	5.47E-07	7.12E-05	zmynd10	ENSDARG00000002406
zgc:165604	9.74196979	3.85319709	0.77359813	4.98087698	6.33E-07	8.18E-05	zgc:165604	ENSDARG00000021241
ap1m3	19.9764498	-2.272251	0.45707391	-4.9712988	6.65E-07	8.54E-05	ap1m3	ENSDARG00000039512
ENSDARG00000086037	63.2303142	1.52300463	0.30810834	4.94308152	7.69E-07	9.81E-05	NA	ENSDARG00000086037

si:dkey-88l16.5	74.4171913	1.07082107	0.21677795	4.93971409	7.82E-07	9.91E-05	si:dkey-88l16.5	ENSDARG00000094850
htra1a	14.0505687	2.17446327	0.44109637	4.92967845	8.24E-07	0.00010367	htra1a	ENSDARG00000032831
epha7	103.968279	1.25163522	0.25470059	4.91414335	8.92E-07	0.00011151	epha7	ENSDARG00000004635
ovol1b	287.059276	0.76203726	0.1554104	4.90338652	9.42E-07	0.00011704	ovol1b	ENSDARG00000078256
si:ch211-137a8.4	9309.09198	-0.505581	0.10319803	-4.8991338	9.63E-07	0.00011884	si:ch211-137a8.4	ENSDARG00000078748
si:ch211-105c13.3	39.3785671	-2.0033122	0.41023781	-4.8832948	1.04E-06	0.00012799	si:ch211-105c13.3	ENSDARG00000089441
si:dkey-147f3.4	133.978537	0.95589331	0.19626442	4.87043616	1.11E-06	0.00013574	si:dkey-147f3.4	ENSDARG00000071029
LOC110439871	22.7083425	2.17642386	0.45234416	4.81143357	1.50E-06	0.00018154	LOC110439871	ENSDARG00000095914
si:ch211-170d8.2	827.288941	-0.6338747	0.13183314	-4.8081587	1.52E-06	0.00018339	si:ch211-170d8.2	ENSDARG00000094887
ENSDARG00000034273	32.8693404	2.87015846	0.59803616	4.79930584	1.59E-06	0.00019003	NA	ENSDARG00000034273
abcb5	53.8994801	-1.2983318	0.27056651	-4.798568	1.60E-06	0.00019003	abcb5	ENSDARG00000021787
dnaaf1	8.59817574	4.77953428	0.99657664	4.79595253	1.62E-06	0.00019135	dnaaf1	ENSDARG00000012030
si:ch211-236p5.3	37.3358709	1.44855684	0.30276703	4.78439423	1.72E-06	0.00020147	si:ch211-236p5.3	ENSDARG00000086418
dnah9	13.1374016	4.99785378	1.05641372	4.7309626	2.23E-06	0.00026092	dnah9	ENSDARG00000103383
shc1	880.442327	0.60919863	0.12891888	4.72544158	2.30E-06	0.00026651	shc1	ENSDARG00000075437
sorcs2	11.0805431	4.62010629	0.97837029	4.72224712	2.33E-06	0.00026883	sorcs2	ENSDARG00000077465
abcc6a	61.1571504	1.64162529	0.3477096	4.72125384	2.34E-06	0.00026883	abcc6a	ENSDARG00000016750
she	28.6910094	3.1333517	0.66472065	4.7137872	2.43E-06	0.00027724	she	ENSDARG00000087956
ENSDARG00000089210	57.6740101	1.85856338	0.39506906	4.70440123	2.55E-06	0.00028861	NA	ENSDARG00000089210
ednraa	33.1140344	2.4157807	0.51371597	4.702561	2.57E-06	0.00028953	ednraa	ENSDARG00000011876
agfg1a	380.072004	0.91532342	0.19472508	4.70059346	2.59E-06	0.00029064	agfg1a	ENSDARG00000030020
ror1	341.562064	0.96716845	0.20614734	4.69163689	2.71E-06	0.00030192	ror1	ENSDARG00000015176
ENSDARG00000035367	10.7440915	5.62547391	1.20446112	4.67053175	3.00E-06	0.00033275	NA	ENSDARG00000035367
si:dkey-30j16.3	9.58217365	6.05498616	1.30504245	4.63968523	3.49E-06	0.00038429	si:dkey-30j16.3	ENSDARG00000037587
mdka	280.659661	0.76258489	0.16444692	4.63727075	3.53E-06	0.00038661	mdka	ENSDARG00000036036
ttc25	20.7264026	2.83479165	0.61356258	4.62021597	3.83E-06	0.00041279	ttc25	ENSDARG00000058140
capsla	7.0532686	5.61185838	1.21431058	4.6214358	3.81E-06	0.00041279	capsla	ENSDARG00000103521
tbc1d2b	92.6980267	0.92449562	0.20008496	4.62051524	3.83E-06	0.00041279	tbc1d2b	ENSDARG00000061986
ENSDARG00000099162	139.769009	2.764224	0.60188805	4.59258828	4.38E-06	0.00046881	NA	ENSDARG00000099162
si:ch211-212o1.2	74.6432924	1.02620162	0.22379398	4.58547467	4.53E-06	0.0004824	si:ch211-212o1.2	ENSDARG00000011498
ajap1	5.26977412	5.17882319	1.13312595	4.5703862	4.87E-06	0.00051564	ajap1	ENSDARG00000038655

c1qtnf4	17.3796195	2.25293834	0.49307623	4.56914812	4.90E-06	0.00051587	c1qtnf4	ENSDARG00000024299
cfap126	10.041541	3.22213448	0.70664972	4.55973361	5.12E-06	0.00053663	cfap126	ENSDARG00000070868
spag8	5.51936619	5.24793675	1.15438688	4.54608142	5.47E-06	0.00056955	spag8	ENSDARG00000103843
ENSDARG00000086098	63.4064243	1.83531524	0.40424207	4.54013917	5.62E-06	0.0005827	NA	ENSDARG00000086098
pttg1ipb	287.866177	0.84827849	0.18690395	4.53857974	5.66E-06	0.00058391	pttg1ipb	ENSDARG00000040039
LOC103910581	27.7744266	-1.7818756	0.39298287	-4.5342323	5.78E-06	0.00059291	LOC103910581	ENSDARG00000060627
ftr83	42.9530089	-1.3343286	0.29473951	-4.5271453	5.98E-06	0.00060672	ftr83	ENSDARG00000025403
casz1	87.1476029	1.49527768	0.33027615	4.52735586	5.97E-06	0.00060672	casz1	ENSDARG00000037030
cyp24a1	80.2816044	1.81379598	0.40183507	4.51378212	6.37E-06	0.00064151	cyp24a1	ENSDARG00000103277
dock5	84.3455525	1.51526414	0.33574535	4.51313517	6.39E-06	0.00064151	dock5	ENSDARG00000001968
dnmt3bb.1	4307.3603	0.64426336	0.14287683	4.50922208	6.51E-06	0.00065009	dnmt3bb.1	ENSDARG00000036791
ccdc181	9.85296464	4.29256507	0.95244	4.5069139	6.58E-06	0.00065383	ccdc181	ENSDARG00000062021
phldb1b	2186.7397	0.92819639	0.20626136	4.50009826	6.79E-06	0.0006717	phldb1b	ENSDARG00000079378
serpinb14	90.9761252	-1.2392682	0.27646459	-4.4825566	7.38E-06	0.00072567	serpinb14	ENSDARG00000091801
zgc:174938	22.2541412	-2.0150214	0.45080218	-4.4698572	7.83E-06	0.00076623	zgc:174938	ENSDARG00000075622
LOC100535393	7.65457988	5.12411199	1.14691233	4.46774514	7.90E-06	0.00076995	LOC100535393	ENSDARG00000017391
tppp3	20.7051107	2.44805729	0.55242779	4.43145206	9.36E-06	0.00090713	tppp3	ENSDARG00000030463
vcana	1045.09666	0.81961172	0.18537306	4.42141761	9.81E-06	0.00094558	vcana	ENSDARG00000103515
sypl2b	199.482047	0.90261967	0.20444745	4.41492266	1.01E-05	0.00096959	sypl2b	ENSDARG00000000690
dnai2b	7.88440831	4.27109124	0.97093518	4.39894581	1.09E-05	0.00103864	dnai2b	ENSDARG00000074081
micall2b	91.6545804	0.87249934	0.19842295	4.3971695	1.10E-05	0.00104204	micall2b	ENSDARG00000017834
mef2cb	18.0380682	1.80725042	0.41156558	4.39116026	1.13E-05	0.00106604	mef2cb	ENSDARG00000009418
zgc:100864	433.999605	-1.2722426	0.29019115	-4.3841537	1.16E-05	0.00109559	zgc:100864	ENSDARG00000039669
ccdc151	29.4379167	2.60531907	0.59553469	4.37475616	1.22E-05	0.00113834	ccdc151	ENSDARG00000062978
tekt1	8.35988602	4.35952795	0.99862339	4.36553759	1.27E-05	0.00117715	tekt1	ENSDARG00000101331
fbln2	154.073446	1.02337827	0.23443293	4.36533506	1.27E-05	0.00117715	fbln2	ENSDARG00000015156
gna15.1	42.3156955	-1.5629195	0.35942699	-4.3483643	1.37E-05	0.00126596	gna15.1	ENSDARG00000016364
sox13	655.16113	0.80990436	0.18659348	4.34047505	1.42E-05	0.00130606	sox13	ENSDARG00000030297
ENSDARG00000097714	12.8359679	3.39579152	0.78334108	4.33501014	1.46E-05	0.00133261	NA	ENSDARG00000097714
wdr93	6.77441685	3.99829193	0.92679986	4.31408345	1.60E-05	0.00145842	wdr93	ENSDARG00000087191
fosl2	152.8822	1.88497692	0.43750491	4.30847035	1.64E-05	0.00148894	fosl2	ENSDARG00000040623

cep112	29.1931359	1.31893654	0.30743289	4.29016078	1.79E-05	0.00160963	cep112	ENSDARG00000079679
ctnnd1	4245.85597	0.3954281	0.09246032	4.27673303	1.90E-05	0.0017019	ctnnd1	ENSDARG00000078233
ENSDARG00000094488	38.3815985	1.75761889	0.4112946	4.2733819	1.93E-05	0.00171973	NA	ENSDARG00000094488
itga4	68.2158457	1.24818203	0.2933937	4.25429057	2.10E-05	0.00186462	itga4	ENSDARG00000103056
ak8	9.8548136	3.44085504	0.80904139	4.25300248	2.11E-05	0.00186682	ak8	ENSDARG00000030961
nfatc3b	39.2817803	1.3561624	0.31899883	4.25130844	2.13E-05	0.00187244	nfatc3b	ENSDARG00000051729
ccdc24	6.51530193	3.92248504	0.92303322	4.24956001	2.14E-05	0.00187858	ccdc24	ENSDARG00000038793
plxnb2a	461.721889	0.80605401	0.18982532	4.2462935	2.17E-05	0.00189758	plxnb2a	ENSDARG00000003811
fhl2b	6.27059975	5.43834286	1.28702031	4.22552993	2.38E-05	0.00207197	fhl2b	ENSDARG00000003991
si:ch73-335m24.5	10.6409738	2.6178188	0.61983351	4.22342253	2.41E-05	0.00208212	si:ch73-335m24.5	ENSDARG00000036383
tuft1a	15.3905934	2.27844123	0.54030491	4.21695452	2.48E-05	0.0021332	tuft1a	ENSDARG00000061242
rsph9	7.43519488	5.06332181	1.20680453	4.19564368	2.72E-05	0.00233366	rsph9	ENSDARG00000017355
efcab6	10.9732312	3.10416985	0.74013979	4.19403185	2.74E-05	0.00233996	efcab6	ENSDARG00000020735
bcl6aa	38.8871666	2.265646	0.54099132	4.18795257	2.81E-05	0.00238252	bcl6aa	ENSDARG00000070864
LOC100330978	14.6644193	-1.9350462	0.46198685	-4.1885309	2.81E-05	0.00238252	LOC100330978	ENSDARG00000086874
si:ch211-199g17.2	655.441349	0.96358017	0.23040876	4.18204667	2.89E-05	0.00243465	si:ch211-199g17.2	ENSDARG00000092310
pcdh1b	6157.59292	-0.6391348	0.15289561	-4.1802035	2.91E-05	0.00244384	pcdh1b	ENSDARG00000036175
plppr5b	3.51800111	5.18364444	1.24645757	4.15870106	3.20E-05	0.00267404	plppr5b	ENSDARG00000101348
ENSDARG00000098949	4.43049371	4.92792738	1.18710371	4.1512189	3.31E-05	0.00275113	NA	ENSDARG00000098949
fcho2	418.429898	0.49478054	0.11954929	4.13871571	3.49E-05	0.002893	fcho2	ENSDARG00000035389
myod1	33.4517781	-1.4345536	0.34680688	-4.1364623	3.53E-05	0.00290911	myod1	ENSDARG00000030110
nlrc5	34.3483667	1.84405625	0.44630589	4.13182141	3.60E-05	0.00295591	nlrc5	ENSDARG00000024631
wdr78	5.11436931	4.52984461	1.09685675	4.1298416	3.63E-05	0.0029689	wdr78	ENSDARG00000044400
grhl3	35.1841682	-1.6819411	0.4076061	-4.1263885	3.69E-05	0.00300114	grhl3	ENSDARG00000078552
myo10l3	629.382931	0.88052157	0.21466996	4.10174562	4.10E-05	0.00332548	myo10l3	ENSDARG00000074143
cdc42ep4b	312.916949	1.25602455	0.30680473	4.09388909	4.24E-05	0.00342592	cdc42ep4b	ENSDARG00000045036
tgm5l	16.6331846	-2.032966	0.49692933	-4.0910566	4.29E-05	0.00345365	tgm5l	ENSDARG00000098837
lcp1	39.4765939	-1.7856407	0.4367267	-4.0886914	4.34E-05	0.00347463	lcp1	ENSDARG00000023188
pclob	25.2766206	2.27098247	0.55581869	4.08583322	4.39E-05	0.00349858	pclob	ENSDARG00000098880
wdr63	16.6622025	2.18135797	0.53396775	4.0851867	4.40E-05	0.00349858	wdr63	ENSDARG00000105093
selenop2	72.7768561	0.99646575	0.24530334	4.0621776	4.86E-05	0.0038463	selenop2	ENSDARG00000079727

rbms1a	2238.64697	0.49636956	0.12263531	4.04752558	5.18E-05	0.00407846	rbms1a	ENSDARG00000074023
shank2	12.5466232	2.53846557	0.62745309	4.04566593	5.22E-05	0.00409433	shank2	ENSDARG00000062325
zgc:101744	353.340578	0.81299534	0.20125246	4.03967898	5.35E-05	0.00418332	zgc:101744	ENSDARG00000038694
dnase1l4.1	188.066788	-0.7631025	0.1891469	-4.0344437	5.47E-05	0.00426051	dnase1l4.1	ENSDARG00000015123
syt6b	4.24813467	4.84817241	1.20396359	4.02684303	5.65E-05	0.00438294	syt6b	ENSDARG00000031463
hip1rb	58.9910676	-1.4111222	0.35057177	-4.0252021	5.69E-05	0.00439603	hip1rb	ENSDARG00000102458
rab25a	35.6599571	-1.9995215	0.49724721	-4.0211819	5.79E-05	0.004454	rab25a	ENSDARG00000058800
grk5	9.71904384	-2.3036793	0.57316624	-4.0192167	5.84E-05	0.00447356	grk5	ENSDARG00000032801
casc1	7.21028686	3.80081486	0.94891114	4.00544866	6.19E-05	0.00472364	casc1	ENSDARG00000016815
ENSDARG00000088187	16.4946973	2.02586418	0.50662506	3.99874454	6.37E-05	0.00484038	NA	ENSDARG00000088187
prrt2	14.2820062	2.163087	0.542548	3.98690435	6.69E-05	0.00506843	prrt2	ENSDARG00000103588
marco	3.46106019	5.17635876	1.29948484	3.98339295	6.79E-05	0.00512392	marco	ENSDARG00000059294
foxa3	1072.562	0.81097952	0.20381606	3.97897746	6.92E-05	0.00519978	foxa3	ENSDARG00000012788
glula	233.42634	-1.0343574	0.2600959	-3.9768309	6.98E-05	0.0052216	glula	ENSDARG00000099776
cfap43	17.4945042	2.31382197	0.58192605	3.97614432	7.00E-05	0.0052216	cfap43	ENSDARG00000001825
apoa1a	29.827961	-1.7004546	0.43016714	-3.953009	7.72E-05	0.00573132	apoa1a	ENSDARG00000012076
bmp3	15.7033539	2.84395931	0.72180768	3.94005133	8.15E-05	0.00602679	bmp3	ENSDARG00000060526
chac1	931.379578	1.08822709	0.27630305	3.93852719	8.20E-05	0.00604214	chac1	ENSDARG00000070426
wipi1	144.071325	0.98795842	0.25119632	3.93301313	8.39E-05	0.00614382	wipi1	ENSDARG00000040657
scel	26.636314	-1.9547232	0.49714383	-3.9319069	8.43E-05	0.00614382	scel	ENSDARG00000034677
six4b	519.999583	0.78360492	0.19929954	3.931795	8.43E-05	0.00614382	six4b	ENSDARG00000031983
prtfdc1	1003.10856	0.59069353	0.15037645	3.92809854	8.56E-05	0.00621564	prtfdc1	ENSDARG00000011683
cfap45	12.4929732	3.30044731	0.84059185	3.92633752	8.62E-05	0.00623793	cfap45	ENSDARG00000068103
slc1a4	96.8654558	1.03188901	0.26362673	3.91420486	9.07E-05	0.00653565	slc1a4	ENSDARG00000000551
dnah6	12.4367842	2.55540484	0.65310346	3.91271059	9.13E-05	0.00655188	dnah6	ENSDARG00000000606
pi4k2a	431.743325	0.71115052	0.18185783	3.91047504	9.21E-05	0.00658842	pi4k2a	ENSDARG00000033666
zgc:136472	17.6812781	1.8184318	0.46695114	3.89426567	9.85E-05	0.00701897	zgc:136472	ENSDARG00000058445
si:ch211-226m16.2	131.493178	0.76570597	0.19737659	3.87941645	0.00010471	0.00743422	si:ch211-226m16.2	ENSDARG00000036785
acaca	43.2544414	-1.2722337	0.32821123	-3.876265	0.00010607	0.00750364	acaca	ENSDARG00000078512
gsap	29.7193755	1.63263146	0.42282904	3.8612094	0.00011283	0.00795247	gsap	ENSDARG00000045481
ENSDARG00000087070	104.890908	0.8606881	0.22314484	3.85708278	0.00011475	0.00802948	NA	ENSDARG00000087070

dmbx1a	504.571005	0.97917569	0.25382936	3.85761392	0.0001145	0.00802948	dmbx1a	ENSDARG00000009922
si:ch73-63e15.2	102.093762	4.89293341	1.26916443	3.85523995	0.00011562	0.00806111	si:ch73-63e15.2	ENSDARG00000016188
atp2a3	49.7315288	1.18317123	0.30730382	3.8501676	0.00011804	0.00811316	atp2a3	ENSDARG00000060978
hbae3	6.10960299	4.80924353	1.24886926	3.85087831	0.00011769	0.00811316	hbae3	ENSDARG00000079305
efhc1	30.5283186	2.25860105	0.58638337	3.85174815	0.00011728	0.00811316	efhc1	ENSDARG00000009743
si:ch211-103e16.5	13.6381785	2.40379196	0.62414243	3.85135163	0.00011747	0.00811316	si:ch211-103e16.5	ENSDARG00000096081
ttc6	13.2649106	1.91900435	0.49894482	3.84612539	0.00012	0.00821897	ttc6	ENSDARG00000104125
aadacl4	25.0734378	1.75617125	0.45701114	3.8427318	0.00012167	0.00830414	aadacl4	ENSDARG00000044802
trim32	99.9961195	-0.9100163	0.23780035	-3.8268079	0.00012982	0.00882884	trim32	ENSDARG00000102505
tshz3b	114.357673	1.01062737	0.26425675	3.82441468	0.00013108	0.00884443	tshz3b	ENSDARG00000103361
tsga10	19.3898531	1.810524	0.47358408	3.82302547	0.00013182	0.00884443	tsga10	ENSDARG00000052512
ccdc173	9.88453151	3.31859975	0.86807544	3.82293934	0.00013187	0.00884443	ccdc173	ENSDARG00000077928
f3a	16.8427561	2.43472019	0.63665424	3.82424245	0.00013117	0.00884443	f3a	ENSDARG00000099124
phospho1	74.3422673	1.28020601	0.33511952	3.82014758	0.00013337	0.0089143	phospho1	ENSDARG00000008403
slc22a31	399.995954	0.7000158	0.18329922	3.81897863	0.00013401	0.00892586	slc22a31	ENSDARG00000078882
plcd1a	53.0020923	-1.6063044	0.42078739	-3.8173778	0.00013488	0.00895319	plcd1a	ENSDARG00000059123
s1pr5a	485.81066	1.28038106	0.33567881	3.81430409	0.00013657	0.00903441	s1pr5a	ENSDARG00000040526
sema6dl	4615.507	0.42995124	0.11285828	3.80965601	0.00013916	0.00914354	sema6dl	ENSDARG00000011533
si:ch73-347e22.8	28.8878537	-1.669854	0.43824359	-3.8103329	0.00013878	0.00914354	si:ch73-347e22.8	ENSDARG00000103322
atp8a2	62.3333907	1.14128638	0.30026518	3.80092824	0.00014416	0.00943972	atp8a2	ENSDARG00000077492
heyl	7.85503336	4.2921302	1.13004916	3.79818008	0.00014576	0.00951284	heyl	ENSDARG00000055798
cyp26c1	68.0044599	1.55971207	0.41087964	3.79603157	0.00014703	0.00956341	cyp26c1	ENSDARG00000056029
maats1	4.20487135	4.23748229	1.11924881	3.78600563	0.00015309	0.00992411	maats1	ENSDARG00000006292
armc3	17.8603618	2.59072195	0.68524085	3.78074653	0.00015636	0.01010234	armc3	ENSDARG00000074708
umps	1301.45693	-0.4747481	0.12579082	-3.7741077	0.00016058	0.01034069	umps	ENSDARG00000012215

Appendix 5 – Table of primers

RT-qPCR primers		Sequence 5' - 3'	
	Name	Forward	Reverse
	lfl2_fw	GGACATGGGCGCACCAGAACT	TACCCGGCCGGCTCGATGAT
	tbxta	AAGACGCGGAGTTGTGGACC	ACTGGCTCTGAGCACGGGAA
	noto_fw	CTGTTGGCATCTGCTCTCCA	TCTCCATTTGATGCGCCTGT
	foxa2_fw	CGGCCAGTCGAACATAAACA	GCTGGATGGCCATGGTTATT
	dusp4_fw	TATCCCGGTGGAAGACAACC	ACCGTTGGAATCCTTGACGG
	dusp6_fw	CGGCTCCGTGTTGGGTTTA	CCGTCGAGGTTCGTCTCAC
	sox32_fw	TCAGCAAAATACTTGGCAAGACA	GACGGGGCCGGTATTTGTAG
	gata5_fw	ACTCTACCGGGAAGGAGGTC	TAGTGTCCGGTTCCGTCTCT
	eve1	CCTCCAGAAAAGCTTTCTTCT CTAT	CAGAGGGAGGTGTTAAATTGTC TT
	itga6a	TCACA TTCTTCCGGCTTCCT	TGGCAGCTCGTATCTCTCTG
	cxcl12b	TCTGACACCTCACACA TGCA	GCAGA TTTGGGAGTTCAGCC
	sox19b	ATTAAACTCGCACACGAACCT T	CCTGAAA TGAGTGGCTTTTCTT
	foxh1	TCTGCAGTCAAGGTGA TGGT	CAGTCTGAGGGGTTGAGGAG
	foxa2	CCTGTGGCCCAA TTGAAGAG	CCTGCGAGTGACTGCAATAC
	elf2	TGCTGTGCGTGACATGAGGCAG	CCGCAACCTTTGGAACGGTGT
	gfp	AAGCTGACCCTGAAGTTCATCTGC	CTTGTAGTTGCCGTCGTCCTTGAA
	nanog	ACACTATGACGGCTTGCACCG	CCCACTGACTGCACGA TCTGG
	pau5f	TTCAGCCGCTACCCCGACCA	ACCAGGGTGTCGCCCTCGAA
	sox2	GTTCCCTGCAGCACA TTTCACG	GCTGGTGCTTTACACACTCAACCT
	sox32	GGACA TGGGCGCACCAGAACT	TACCCGGCCGGCTCGATGAT
	sox17	TCGCTGGACGTCA TCGCTTG	CTCCGTCTTGAGCCTCGTCG
	gata5	GCAGGAACACGACTGGGGTG	AAGACGCGGAGTTGTGGACC
	mixl1		ZF_SOX2_1
	crcx4		ZF_FOXI1_2
	foxa2		ZF_TBX24_1
	tbxta		ZF_IRX7_1
	myf5		ZF_TFAP2A_1
	vox		ZF_VOX_2
	tbx16		ZF_BMP4_1
	tbx24		ZF_TBX16_1
	bmb4		ZF_CHD_1
	tfap2b		ZF_FRZB_1
	otx2	ACTGGCTCTGAGCACGGGAA	AATGCATACCGGTGCGAGGG
	foxi1	AGCAGAACCCAGATCTGCAC	GCTTCTCTGCCAAGGTCAAC
	sox2	CGGGATGAAAACGTCCATTT	ATGACCAGGATCACCAATCCA
	irx7	GAGCCGTGAAGATGGAAGG	TCATGTTGCTCACGGAAGAG

RT-PCR and ISH primers	Name	Sequence 5' - 3' Forward	Reverse
	sox32 HRM 1	CCAGCATACCATTGACTATCCTAAC	5'-CCACTTGATGATGTTGCCTCG-3'
	sox32 HRM 2	CCAGCATACCATTGACTATCCTAAC	5'-CCACTTGATGATGTTGCCTCG-3'
	sox32 HRM 3	GCATAAATCCCAAACAAAAGCC	5'-ACG TCA GCT CTC CAA ATG-3'
	mixl1 HRM 1	ACGTCAGCTCTCCAAATGCC	5'-TGTGGGGAAGCCTATATGAGTT-3'
	mixl1 HRM 2	GCCCCTAGCACAATGAAGAT	GTTTGAGTCGGCGTGAAGT
	txn	AGTTGGTGGTGGTGGACTTC	TAATACGACTCACTATAGGTTTTCATTTCATACAAAGCCAACA
	prex1	TTGATGCAGACCATGGACCTCAT	TAATACGACTCACTATAGAAACTCACCAGCTCCCACTG
	prdx5	CCAGACTGCCCAATCGTAAT	TAATACGACTCACTATAGGTGCGGTTTACAGCCAAACT
	pck2	TAAAGGCGGCATCTGGTCCATT	TAATACGACTCACTATAGCCACTGTCGTCACCTGCTTA
	gfp	CTACCTGTTCCATGGCCAAC	TAATACGACTCACTATAGAAAGGGCAGATTGTGTGGAC
	cdh6	CACCGATATCAACGACAACG	TAATACGACTCACTATAGCTCGATCCAGAGGTTTCTGC

ChIP-PCR primers	Name	Sequence 5' - 3' Forward	Reverse
	negative region	GACTCCACACAA TCTGCAACA T	ACCACCTACGCTAAAGAAACCA
	sox32 peak 1	GCCCGTGTGTAGTGAGAGAT	AACTGCGCGACTATTTCTGA
	sox32 peak 2	AA TTGTGCGAAA TGGCCAC	ACTACAGCACAGTCACACGT
	sox17 peak	TAAGCCGTGTAACCATGCAG	TGA TGTTGCGACCTGTGAAC
	gata4 peak 1	TCC CGG GTC ATC GCA CAT CAA G	TCAAGAGACCCATGAACGCC
	gata4 peak 2	TCC CCT CAG CTG TTT TGA CTC TGC 3	TCCGGGTGTTCCTTCATGTG
	gata6 peak	TGG TCG CCA ATC AGT CTC CTC G	GAACGCGTTTATGGTGTGGG
	foxa2 peak 1	CGC GAG CAA TGA GTT CAC AGG TC	ACATCCCAGGAAGCGGAAAG
	foxa2 peak 2	GACTCCACACAATCTGCAACAT	TCTGGGACTTCCTTGTGGTC
	foxa2 peak 3	ACCACCTACGCTAAAGAAACCA	GAAAGTCGGTGGGATTCAGA
	foxa3 peak 1	ATG GGC AGT TCA GGT ACG CAG G	AAAGAGCTGGAGGTGAAGGC
	foxa3 peak 2	TGC ACG CTC CCA TCA ATG CAC	GACTCGGGACTCAAAGCTGG
	vox peak	AGGAGAGTGACATTGGCAGC	GGTCGCCTGTCTCTTCTCAG
	vent Peak 1	CCGTACATGCAAGAAGCAGA	A TCAAAGGTGGCA TTTGGAG
	vent Peak 2	GCCCGTGTGTAGTGAGAGAT	AACTGCGCGACTATTTCTGA
	gsc peak	AA TTGTGCGAAA TGGCCAC	ACTACAGCACAGTCACACGT
	sebox peak	ACCCAAAGGAACACGAACAG	AACAGGCGATGTGTTTAGGG
	dusp6 Peak1	TAAGCCGTGTAACCATGCAG	TGA TGTTGCGACCTGTGAAC
	dusp6 peak 2	AGAGACGGAACCGGACACTA	TCTCCTCCACAGTGTTGTCG
	pcdh8 peak	TCTCACTATGGGCACAGCAG	GGGACAGCTTCAGAGCAGAC
	prmd1a peak	AGTTGCTGGAGCTCGTCTTC	TCGGTTCTGGAACCACAC
	irx3a peak	GAAATCCACCAAAGCTCACG	TGTAGGCAGTACGGGTCCTC
	mixl1 peak	ACGCCAAGGTCTGAAGGTC	AGCGTGGCTTCTTCTACACAC
	gata5 peak 1	CGTTTTCCCACTCATCTCTG	GATCTGTACCGCAGGCACTC
	gata5 peak 2	GGAGACGAAGTACCCAGACG	GCTTCTGTCGTCTCCACTTG
	dusp4 peak	AAGCTGACCCTGAAGTTCATCTGC	CTTGTAGTTGCCGTCGTCCTTGAA
	tbxta peak	TCACA TTCTTCCGGCTTCCT	TGGCAGCTCGTATCTCTCTG
	dlc peak	TCTGACACCTCACACA TGCA	GCAGA TTTGGGAGTTCAGCC
	txn peak	ATTAAACTCGCACACGAACCT T	CCTGAAA TGAGTGGCTTTTCTT
	prdx5 peak 1	TCTGCAGTCAAGGTGA TGGT	CAGTCTGAGGGGTTGAGGAG
	prdx5 peak 2	CCTGTGGCCCAA TTGAAGAG	CCTGCGAGTGACTGCAATAC
	cdx4 peak	GACTCCACACAA TCTGCAACA T	ACCACCTACGCTAAAGAAACCA
	nanog peak	ACCCAAAGGAACACGAACAG	AACAGGCGATGTGTTTAGGG

Bibliography

Aamar, E., Dawid, I.B., 2010. Sox17 and chordin are required for formation of Kupffer's vesicle and left-right asymmetry determination in zebrafish. Developmental Dynamics 239, 2980-2988.

Aanes, H., Winata, C.L., Lin, C.H., Chen, J.P., Srinivasan, K.G., Lee, S.G., Lim, A.Y., Hajan, H.S., Collas, P., Bourque, G., Gong, Z., Korzh, V., Alestrom, P., Mathavan, S., 2011. Zebrafish mRNA sequencing deciphers novelties in transcriptome dynamics during maternal to zygotic transition. Genome Res 21, 1328-1338.

Aday, A.W., Zhu, L.J., Lakshmanan, A., Wang, J., Lawson, N.D., 2011. Identification of cis regulatory features in the embryonic zebrafish genome through large-scale profiling of H3K4me1 and H3K4me3 binding sites. Developmental Biology 357, 450-462.

Afouda, B.A., Ciau-Uitz, A., Patient, R., 2005. GATA4, 5 and 6 mediate TGFbeta maintenance of endodermal gene expression in Xenopus embryos. Development 132, 763.

Agarwal, P., Verzi, M.P., Nguyen, T., Hu, J., Ehlers, M.L., McCulley, D.J., Xu, S.M., Dodou, E., Anderson, J.P., Wei, M.L., Black, B.L., 2011. The MADS box transcription factor MEF2C regulates melanocyte development and is a direct transcriptional target and partner of SOX10. Development 138, 2555-2565.

Ahnert, S.E., Fink, T.M.A., 2016. Form and function in gene regulatory networks: the structure of network motifs determines fundamental properties of their dynamical state space. Journal of the Royal Society, Interface 13, 20160179.

Alexa, K., Choe, S.-K., Hirsch, N., Etheridge, L., Laver, E., Sagerström, C.G., 2009. Maternal and Zygotic aldh1a2 Activity Is Required for Pancreas Development in Zebrafish. PLOS ONE 4, e8261.

Alexander, J., Rothenberg, M., Henry, G.L., Stainier, D.Y., 1999. Casanova plays an early and essential role in endoderm formation in zebrafish. Dev Biol 215.

Alexander, J., Stainier, D.Y., 1999. A molecular pathway leading to endoderm formation in zebrafish. Curr Biol 9.

Altman, N., Krzywinski, M., 2015. Points of significance: Sources of variation. Nat Methods 12, 5-6.

Anders, S., Pyl, P.T., Huber, W., 2015. HTSeq—a Python framework to work with high-throughput sequencing data. Bioinformatics 31, 166-169.

Anderson, R.M., Delous, M., Bosch, J.A., Ye, L., Robertson, M.A., Hesselson, D., Stainier, D.Y.R., 2013. Hepatocyte Growth Factor Signaling in Intrapancreatic Ductal Cells Drives Pancreatic Morphogenesis. PLOS Genetics 9, e1003650.

Andersson, Baechi, Hoechl, Richter, 1998. Autofluorescence of living cells. 191, 1-7.

Andrews, S., 2010. FastQC A Quality Control tool for High Throughput Sequence Data.

Ang, L.T., Tan, A.K.Y., Autio, M.I., Goh, S.H., Choo, S.H., Lee, K.L., Tan, J., Pan, B., Lee, J.J.H., Lum, J.J., Lim, C.Y.Y., Yeo, I.K.X., Wong, C.J.Y., Liu, M., Oh, J.L.L., Chia, C.P.L., Loh, C.H., Chen, A., Chen, Q., Weissman, I.L., Loh, K.M., Lim, B., 2018. A Roadmap for Human Liver Differentiation from Pluripotent Stem Cells. Cell Reports 22, 2190-2205.

Aoki, T.O., David, N.B., Minchiotti, G., Saint-Etienne, L., Dickmeis, T., Persico, G.M., Strähle, U., Mourrain, P., Rosa, F.M., 2002a. Molecular integration of casanova in the Nodal signalling pathway controlling endoderm formation. Development 129, 275-286.

Aoki, T.O., Mathieu, J., Saint-Etienne, L., Rebagliati, M.R., Peyrieras, N., Rosa, F.M., 2002b. Regulation of nodal signalling and mesendoderm formation by TARAM-A, a TGFbeta-related type I receptor. Dev Biol 241, 273-288.

Argenton, F., Zecchin, E., Bortolussi, M., 1999. Early appearance of pancreatic hormone-expressing cells in the zebrafish embryo. Mechanisms of development 87, 217-221.

Ari, Ş., Arikan, M., 2016. Next-Generation Sequencing: Advantages, Disadvantages, and Future, in: Hakeem, K.R., Tombuloğlu, H., Tombuloğlu, G. (Eds.), Plant Omics: Trends and Applications. Springer International Publishing, Cham, pp. 109-135.

Artus, J., Piliszek, A., Hadjantonakis, A.K., 2011. The primitive endoderm lineage of the mouse blastocyst: sequential transcription factor activation and regulation of differentiation by Sox17. Dev Biol 350, 393-404.

Azevedo, A.S., Grotek, B., Jacinto, A., Weidinger, G., Saúde, L., 2011. The Regenerative Capacity of the Zebrafish Caudal Fin Is Not Affected by Repeated Amputations. PLOS ONE 6, e22820.

Azpeitia, E., Weinstein, N., Benítez, M., Mendoza, L., Alvarez-Buylla, E., 2013. Finding Missing Interactions of the Arabidopsis thaliana Root Stem Cell Niche Gene Regulatory Network. 4.

Babb, S.G., Barnett, J., Doedens, A.L., Cobb, N., Liu, Q., Sorkin, B.C., Yelick, P.C., Raymond, P.A., Marrs, J.A., 2001. Zebrafish E-cadherin: expression during early embryogenesis and regulation during brain development. Dev Dyn 221, 231-237.

Babb, S.G., Marrs, J.A., 2004. E-cadherin regulates cell movements and tissue formation in early zebrafish embryos. Developmental Dynamics 230, 263-277.

Baedke, J., 2013. The epigenetic landscape in the course of time: Conrad Hal Waddington's methodological impact on the life sciences. Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences 44, 756-773.

Bagatto, B., Francl, J., Liu, B., Liu, Q., 2006. Cadherin2 (N-cadherin) plays an essential role in zebrafish cardiovascular development. BMC developmental biology 6, 23-23.

Baranello, L., Kouzine, F., Sanford, S., Levens, D., 2016. ChIP bias as a function of cross-linking time. Chromosome research : an international journal on the molecular, supramolecular and evolutionary aspects of chromosome biology 24, 175-181.

Baruzzo, G., Hayer, K.E., Kim, E.J., Di Camillo, B., FitzGerald, G.A., Grant, G.R., 2017. Simulation-based comprehensive benchmarking of RNA-seq aligners. Nature methods 14, 135-139.

Bazzini, A.A., Lee, M.T., Giraldez, A.J., 2012. Ribosome profiling shows that miR-430 reduces translation before causing mRNA decay in zebrafish. Science 336, 233-237.

Bedell, V.M., Westcot, S.E., Ekker, S.C., 2011. Lessons from morpholino-based screening in zebrafish. Briefings in functional genomics 10, 181-188.

Bennett, J.T., Joubin, K., Cheng, S., Aanstad, P., Herwig, R., Clark, M., Lehrach, H., Schier, A.F., 2007. Nodal Signaling Activates Differentiation Genes During Zebrafish Gastrulation. Developmental biology 304, 525-540.

Bernard, P., Tang, P., Liu, S., Dewing, P., Harley, V.R., Vilain, E., 2003. Dimerization of SOX9 is required for chondrogenesis, but not for sex determination. Human Molecular Genetics 12, 1755-1765.

Bhatia, S., Monahan, J., Ravi, V., Gautier, P., Murdoch, E., Brenner, S., van Heyningen, V., Venkatesh, B., Kleinjan, D.A., 2014. A survey of ancient conserved non-coding elements in the PAX6 locus reveals a landscape of interdigitated cis-regulatory archipelagos. Developmental Biology 387, 214-228.

Biehlmaier, O., Makhankov, Y., Neuhauss, S.C.F., 2007. Impaired Retinal Differentiation and Maintenance in Zebrafish Laminin Mutants. Investigative Ophthalmology & Visual Science 48, 2887-2894.

Birnbaum, R.Y., Patwardhan, R.P., Kim, M.J., Findlay, G.M., Martin, B., Zhao, J., Bell, R.J.A., Smith, R.P., Ku, A.A., Shendure, J., Ahituv, N., 2014. Systematic dissection of coding exons at single nucleotide resolution supports an additional role in cell-specific transcriptional regulation. PLoS genetics 10, e1004592-e1004592.

Bischof, J., Driever, W., 2004. Regulation of hhex expression in the yolk syncytial layer, the potential Nieuwkoop center homolog in zebrafish. Dev Biol 276, 552-562.

Bisgrove, B.W., Su, Y.-C., Yost, H.J., 2017. Maternal Gdf3 is an obligatory cofactor in Nodal signaling for embryonic axis formation in zebrafish. eLife 6, e28534.

Bjornson, C.R., Griffin, K.J., Farr, G.H., Terashima, A., Himeda, C., Kikuchi, Y., Kimelman, D., 2005. Eomesodermin is a localized maternal determinant required for endoderm induction in zebrafish. Dev Cell 9.

Bogdanovic, O., Fernandez-Minan, A., Tena, J.J., de la Calle-Mustienes, E., Gomez-Skarmeta, J.L., 2013. The developmental epigenomics toolbox: ChIP-seq and MethylCap-seq profiling of early zebrafish embryos. Methods 62, 207-215.

Bolouri, H., Davidson, E.H., 2003. Transcriptional regulatory cascades in development: Initial rates, not steady state, determine network kinetics. 100, 9371-9376.

Botstein, D., White, R.L., Skolnick, M., Davis, R.W., 1980. Construction of a genetic linkage map in man using restriction fragment length polymorphisms. American journal of human genetics 32, 314-331.

Bowles, J., Schepers, G., Koopman, P., 2000. Phylogeny of the SOX Family of Developmental Transcription Factors Based on Sequence and Structural Indicators. Developmental Biology 227, 239-255.

Briggs, J.A., Weinreb, C., Wagner, D.E., Megason, S., Peshkin, L., Kirschner, M.W., Klein, A.M., 2018. The dynamics of gene expression in vertebrate embryogenesis at single-cell resolution. 360, eaar5780.

Bronner, G., Chu-LaGraff, Q., Doe, C.Q., Cohen, B., Weigel, D., Taubert, H., Jackle, H., 1994. Sp1/egr-like zinc-finger protein required for endoderm specification and germ-layer formation in Drosophila. Nature 369, 664-668.

Brown, J.L., Snir, M., Noushmehr, H., Kirby, M., Hong, S.K., Elkahloun, A.G., Feldman, B., 2008. Transcriptional profiling of endogenous germ layer precursor cells identifies dusp4 as an essential gene in zebrafish endoderm specification. Proc Natl Acad Sci U S A 105, 12337-12342.

Bruce, A.E., Howley, C., Dixon Fox, M., Ho, R.K., 2005. T-box gene eomesodermin and the homeobox-containing Mix/Bix gene mtx2 regulate epiboly movements in the zebrafish. Dev Dyn 233.

Buermans, H.P.J., den Dunnen, J.T., 2014. Next generation sequencing technology: Advances and applications. Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease 1842, 1932-1941.

Byron, S.A., Van Keuren-Jensen, K.R., Engelthaler, D.M., Carpten, J.D., Craig, D.W., 2016. Translating RNA sequencing into clinical diagnostics: opportunities and challenges. Nature Reviews Genetics 17, 257.

Carrasco, M., Delgado, I., Soria, B., Martin, F., Rojas, A., 2012. GATA4 and GATA6 control mouse pancreas organogenesis. J Clin Invest 122, 3504-3515.

Carroll, T.S., Liang, Z., Salama, R., Stark, R., de Santiago, I., 2014. Impact of artifact removal on ChIP quality metrics in ChIP-seq and ChIP-exo data. Frontiers in Genetics 5, 75.

Castro, D.M., de Veaux, N.R., Miraldi, E.R., Bonneau, R., 2019. Multi-study inference of regulatory networks for more accurate models of gene regulation. PLOS Computational Biology 15, e1006591.

Cermenati, S., Moleri, S., Cimbro, S., Corti, P., Del Giacco, L., Amodeo, R., Dejana, E., Koopman, P., Cotelli, F., Beltrame, M., 2008. Sox18 and Sox7 play redundant roles in vascular development. Blood 111, 2657-2666.

Chae, S., Lee, H.-K., Kim, Y.-K., Jung Sim, H., Ji, Y., Kim, C., Ismail, T., Park, J.-W., Kwon, O.-S., Kang, B.-S., Lee, D.-S., Bae, J.-S., Kim, S.-H., Min, K.-J., Kyu Kwon, T., Park, M.-J., Han, J.-K., Kwon, T., Park, T.-J., Lee, H.-S., 2017. Peroxiredoxin1, a novel regulator of pronephros development, influences retinoic acid and Wnt signaling by controlling ROS levels. Scientific Reports 7, 8874.

Chan, S.S.-K., Kyba, M., 2013. What is a Master Regulator? Journal of stem cell research & therapy 3, 114.

Chan, T.-M., Chao, C.-H., Wang, H.-D., Yu, Y.-J., Yuh, C.-H., 2009a. Functional analysis of the evolutionarily conserved cis-regulatory elements on the sox17 gene in zebrafish. Developmental Biology 326, 456-470.

Chan, T.-M., Longabaugh, W., Bolouri, H., Chen, H.-L., Tseng, W.-F., Chao, C.-H., Jang, T.-H., Lin, Y.-I., Hung, S.-C., Wang, H.-D., Yuh, C.-H., 2009b. Developmental gene regulatory networks in the zebrafish embryo. Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms 1789, 279-298.

Charney, R.M., Forouzmand, E., Cho, J.S., Cheung, J., Paraiso, K.D., Yasuoka, Y., Takahashi, S., Taira, M., Blitz, I.L., Xie, X., Cho, K.W.Y., 2017a. Foxh1 Occupies cis-Regulatory Modules Prior to Dynamic Transcription Factor Interactions Controlling the Mesendoderm Gene Program. Developmental cell 40, 595-607.e594.

Charney, R.M., Paraiso, K.D., Blitz, I.L., Cho, K.W.Y., 2017b. A gene regulatory program controlling early Xenopus mesendoderm formation: Network conservation and motifs. Semin Cell Dev Biol 66, 12-24.

Chen, J.N., Haffter, P., Odenthal, J., Vogelsang, E., Brand, M., van Eeden, F.J., Furutani-Seiki, M., Granato, M., Hammerschmidt, M., Heisenberg, C.P., Jiang, Y.J., Kane, D.A., Kelsh, R.N., Mullins, M.C., Nusslein-Volhard, C., 1996. Mutations affecting the cardiovascular system and other internal organs in zebrafish. Development 123, 293-302.

Chen, Y., Li, Y., Narayan, R., Subramanian, A., Xie, X., 2016. Gene expression inference with deep learning. Bioinformatics 32, 1832-1839.

Chen, Y., Schier, A.F., 2001. The zebrafish Nodal signal Squint functions as a morphogen. Nature 411.

Chen, Y., Schier, A.F., 2002. Lefty proteins are long-range inhibitors of squint-mediated nodal signaling. Curr Biol 12, 2124-2128.

Cheng, X., Tiyaboonchai, A., Gadue, P., 2013. Endodermal stem cell populations derived from pluripotent stem cells. Current opinion in cell biology 25, 265-271.

Chia, C.Y., Madrigal, P., Denil, S., Martinez, I., Garcia-Bernardo, J., El-Khairi, R., Chhatriwala, M., Shepherd, M.H., Hattersley, A.T., Dunn, N.R., Vallier, L., 2019. GATA6 Cooperates with EOMES/SMAD2/3 to Deploy the Gene Regulatory Network Governing Human Definitive Endoderm and Pancreas Formation. Stem Cell Reports 12, 57-70.

Chocron, S., Verhoeven, M.C., Rentzsch, F., Hammerschmidt, M., Bakkers, J., 2007. Zebrafish Bmp4 regulates left–right asymmetry at two distinct developmental time points. Developmental Biology 305, 577-588.

Cholley, P.-E., Moehlin, J., Rohmer, A., Zilliox, V., Nicaise, S., Gronemeyer, H., Mendoza-Parra, M.A., 2018. Modeling gene-regulatory networks to describe cell fate transitions and predict master regulators. npj Systems Biology and Applications 4, 29.

Chudakov, D.M., Matz, M.V., Lukyanov, S., Lukyanov, K.A., 2010. Fluorescent Proteins and Their Applications in Imaging Living Cells and Tissues. 90, 1103-1163.

Chung, M.I.S., Ma, A.C.H., Fung, T.-K., Leung, A.Y.H., 2011. Characterization of Sry-related HMG box group F genes in zebrafish hematopoiesis. Experimental Hematology 39, 986-998.e985.

Chung, W.-S., Stainier, D.Y.R., 2008. Intra-endodermal interactions are required for pancreatic beta cell induction. Developmental Cell 14, 582-593.

Clay, M.R., Halloran, M.C., 2014. Cadherin 6 promotes neural crest cell detachment via F-actin regulation and influences active Rho distribution during epithelial-to-mesenchymal transition. Development 141, 2506.

Clements, D., Cameleyre, I., Woodland, H.R., 2003. Redundant early and overlapping larval roles of Xsox17 subgroup genes in Xenopus endoderm development. Mechanisms of development 120, 337-348.

Clements, D., Woodland, H.R., 2000. Changes in embryonic cell fate produced by expression of an endodermal transcription factor, Xsox17. Mechanisms of development 99, 65-70.

Collins, M.M., Maischein, H.M., Dufourcq, P., Charpentier, M., Blader, P., Stainier, D.Y., 2018. Pitx2c orchestrates embryonic axis extension via mesendodermal cell migration. Elife 7.

Conesa, A., Madrigal, P., Tarazona, S., Gomez-Cabrero, D., Cervera, A., McPherson, A., Szcześniak, M.W., Gaffney, D.J., Elo, L.L., Zhang, X., Mortazavi, A., 2016. A survey of best practices for RNA-seq data analysis. Genome Biology 17, 1-19.

Consortium, S.M.-I., Su, Z., Łabaj, P.P., Li, S., Thierry-Mieg, J., Thierry-Mieg, D., Shi, W., Wang, C., Schroth, G.P., Setterquist, R.A., Thompson, J.F., Jones, W.D., Xiao, W., Xu, W., Jensen, R.V., Kelly, R., Xu, J., Conesa, A., Furlanello, C., Gao, H., Hong, H., Jafari, N., Letovsky, S., Liao, Y., Lu, F., Oakeley, E.J., Peng, Z., Praul, C.A., Santoyo-Lopez, J., Scherer, A., Shi, T., Smyth, G.K., Staedtler, F., Sykacek, P., Tan, X.-X., Thompson, E.A., Vandesompele, J., Wang, M.D., Wang, J., Wolfinger, R.D., Zavadil, J., Auerbach, S.S., Bao, W., Binder, H., Blomquist, T., Brilliant, M.H., Bushel, P.R., Cai, W., Catalano, J.G., Chang, C.-W., Chen, T., Chen, G., Chen, R., Chierici, M., Chu, T.-M., Clevert, D.-A., Deng, Y., Derti, A., Devanarayan, V., Dong, Z., Dopazo, J., Du, T., Fang, H., Fang, Y., Fasold, M., Fernandez, A., Fischer, M., Furió-Tari, P., Fuscoe, J.C., Caimet, F., Gaj, S., Gandara, J., Gao, H., Ge, W., Gondo, Y., Gong, B., Gong, M., Gong, Z., Green, B., Guo, C., Guo, L., Guo, L.-W., Hadfield, J., Hellemans, J., Hochreiter, S., Jia, M., Jian, M., Johnson, C.D., Kay, S., Kleinjans, J., Lababidi, S., Levy, S., Li, Q.-Z., Li, L., Li, L., Li, P., Li, Y., Li, H., Li, J., Li, S., Lin, S.M., López, F.J., Lu, X., Luo, H., Ma, X., Meehan, J., Megherbi, D.B., Mei, N., Mu, B., Ning, B., Pandey, A., Pérez-Florido, J., Perkins, R.G., Peters, R., Phan, J.H., Pirooznia, M., Qian, F., Qing, T., Rainbow, L., Rocca-Serra, P., Sambourg, L., Sansone, S.-A., Schwartz, S., Shah, R., Shen, J., Smith, T.M., Stegle, O., Stralis-Pavese, N., Stupka, E., Suzuki, Y., Szkotnicki, L.T., Tinning, M., Tu, B., van Delft, J., Vela-Boza, A., Venturini, E., Walker, S.J., Wan, L., Wang, W., Wang, J., Wang, J., Wieben, E.D., Willey, J.C., Wu, P.-Y., Xuan, J., Yang, Y., Ye, Z., Yin, Y., Yu, Y., Yuan, Y.-C., Zhang, J., Zhang, K.K., Zhang, W., Zhang, W., Zhang, Y., Zhao, C., Zheng, Y., Zhou, Y., Zumbo, P., Tong, W., Kreil, D.P., Mason, C.E., Shi, L., 2014. A comprehensive assessment of RNA-seq accuracy, reproducibility and information content by the Sequencing Quality Control Consortium. Nature Biotechnology 32, 903.

Corish, P., Tyler-Smith, C., 1999. Attenuation of green fluorescent protein half-life in mammalian cells. Protein Engineering, Design and Selection 12, 1035-1040.

Cossarizza, A., Chang, H.D., Radbruch, A., Akdis, M., Andra, I., Annunziato, F., Bacher, P., Barnaba, V., Battistini, L., Bauer, W.M., Baumgart, S., Becher, B., Beisker, W., Berek, C., Blanco, A., Borsellino, G., Boulais, P.E., Brinkman, R.R., Buscher, M., Busch, D.H., Bushnell, T.P., Cao, X., Cavani, A., Chattopadhyay, P.K., Cheng, Q., Chow, S., Clerici, M., Cooke, A., Cosma, A., Cosmi, L., Cumano, A., Dang, V.D., Davies, D., De Biasi, S., Del Zotto, G., Della Bella, S., Dellabona, P., Deniz, G., Dessing, M., Diefenbach, A., Di Santo, J., Dieli, F., Dolf, A., Donnenberg, V.S., Dorner, T., Ehrhardt, G.R.A., Endl, E., Engel, P., Engelhardt, B., Esser, C., Everts, B., Dreher, A., Falk, C.S., Fehniger, T.A., Filby, A., Fillatreau, S., Follo, M., Forster, I., Foster, J., Foulds, G.A., Frenette, P.S., Galbraith, D., Garbi, N., Garcia-Godoy, M.D., Geginat, J., Ghoreschi, K., Gibellini, L., Goettlinger, C., Goodyear, C.S., Gori, A., Grogan, J., Gross, M., Grutzkau, A., Grummitt, D., Hahn, J., Hammer, Q., Hauser, A.E., Haviland, D.L., Hedley, D., Herrera, G., Herrmann, M., Hiepe, F., Holland, T., Hombrink, P., Houston, J.P., Hoyer, B.F., Huang, B., Hunter, C.A., Iannone, A., Jack, H.M., Javega, B., Jonjic, S., Juelke, K., Jung, S., Kaiser, T., Kalina, T., Keller, B., Khan, S., Kienhofer, D., Kroneis, T., Kunkel, D., Kurts, C., Kvistborg, P., Lannigan, J., Lantz, O., Larbi, A., LeibundGut-Landmann, S., Leipold, M.D., Levings, M.K., Litwin, V., Liu, Y., Lohoff, M., Lombardi, G., Lopez, L., Lovett-Racke, A., Lubberts, E., Ludewig, B., Lugli, E., Maecker, H.T., Martrus, G., Matarese, G., Maueroder, C., McGrath, M., McInnes, I., Mei, H.E., Melchers, F., Melzer, S., Mielenz, D., Mills, K., Mirrer, D., Mjosberg, J., Moore, J., Moran, B., Moretta, A., Moretta, L., Mosmann, T.R., Muller, S., Muller, W., Munz, C., Multhoff, G., Munoz, L.E., Murphy, K.M., Nakayama, T., Nasi, M., Neudorfl, C., Nolan, J., Nourshargh, S., O'Connor, J.E., Ouyang, W., Oxenius, A., Palankar, R., Panse, I., Peterson, P., Peth, C., Petriz, J., Philips, D., Pickl, W., Piconese, S., Pinti, M., Pockley, A.G., Podolska, M.J., Pucillo, C., Quataert, S.A., Radstake, T., Rajwa, B., Rebhahn, J.A., Recktenwald, D., Remmerswaal, E.B.M., Rezvani, K., Rico, L.G., Robinson, J.P., Romagnani, C., Rubartelli, A., Ruckert, B., Ruland, J., Sakaguchi, S., Sala-de-Oyanguren, F., Samstag, Y., Sanderson, S., Sawitzki, B., Scheffold, A., Schiemann, M., Schildberg, F., Schimisky, E., Schmid, S.A., Schmitt, S., Schober, K., Schuler, T., Schulz, A.R., Schumacher, T., Scotta, C., Shankey, T.V., Shemer, A., Simon, A.K., Spidlen, J., Stall, A.M., Stark, R., Stehle, C., Stein, M., Steinmetz, T., Stockinger, H., Takahama, Y., Tarnok, A., Tian, Z., Toldi, G., Tornack, J., Traggiai, E., Trotter, J., Ulrich, H., van der Braber, M., van Lier, R.A.W., Veldhoen, M., Vento-Asturias, S., Vieira, P., Voehringer, D., Volk, H.D., von Volkmann, K., Waisman, A., Walker, R., Ward, M.D., Warnatz, K., Warth, S., Watson, J.V., Watzl, C., Wegener, L., Wiedemann, A., Wienands, J., Willimsky, G., Wing, J., Wurst, P., Yu, L., Yue, A., Zhang, Q., Zhao, Y., Ziegler, S., Zimmermann, J., 2017. Guidelines for the use of flow cytometry and cell sorting in immunological studies. European journal of immunology 47, 1584-1797.

Costa-Silva, J., Domingues, D., Lopes, F.M., 2017. RNA-Seq differential expression analysis: An extended review and a software tool. PLOS ONE 12, e0190152.

Crews, S.T., Pearson, J.C., 2009. Transcriptional autoregulation in development. Current biology : CB 19, R241-R246.

Dalgin, G., Ward, A.B., Hao, L.T., Beattie, C.E., Nechiporuk, A., Prince, V.E., 2011. Zebrafish mnx1 controls cell fate choice in the developing endocrine pancreas. Development 138, 4597.

David, N.B., Rosa, F.M., 2001. Cell autonomous commitment to an endodermal fate and behaviour by activation of Nodal signalling. Development 128, 3937.

Davidson, E.H., 2009. Network design principles from the sea urchin embryo. Current Opinion in Genetics & Development 19, 535-540.

Davidson, E.H., Rast, J.P., Oliveri, P., Ransick, A., Calestani, C., Yuh, C.-H., Minokawa, T., Amore, G., Hinman, V., Arenas-Mena, C., Otim, O., Brown, C.T., Livi, C.B., Lee, P.Y., Revilla, R., Rust, A.G., Pan, Z.j., Schilstra, M.J., Clarke, P.J.C., Arnone, M.I., Rowen, L., Cameron, R.A., McClay, D.R., Hood, L., Bolouri, H., 2002. A Genomic Regulatory Network for Development. 295, 1669-1678.

Davis, T.L., Rebay, I., 2017. Master regulators in development: Views from the Drosophila retinal determination and mammalian pluripotency gene networks. Developmental Biology 421, 93-107.

de Jong, M., Rauwerda, H., Bruning, O., Verkooijen, J., Spaink, H.P., Breit, T.M., 2010. RNA isolation method for single embryo transcriptome analysis in zebrafish. BMC Research Notes 3, 73.

Decker, K., Goldman, D.C., Grasch, C.L., Sussel, L., 2006. Gata6 is an important regulator of mouse pancreas development. Dev Biol 298, 415-429.

Deplancke, B., Alpern, D., Gardeux, V., 2016. The Genetics of Transcription Factor DNA Binding Variation. Cell 166, 538-554.

Dey, B., Thukral, S., Krishnan, S., Chakrobarty, M., Gupta, S., Manghani, C., Rani, V., 2012. DNA–protein interactions: methods for detection and analysis. Molecular and Cellular Biochemistry 365, 279-299.

Dick, A., Mayr, T., Bauer, H., Meier, A., Hammerschmidt, M., 2000. Cloning and characterization of zebrafish smad2, smad3 and smad4. Gene 246, 69-80.

Dickmeis, T., Mourrain, P., Saint-Etienne, L., Fischer, N., Aanstad, P., Clark, M., Strahle, U., Rosa, F., 2001. A crucial component of the endoderm formation pathway, CASANOVA, is encoded by a novel sox-related gene. Genes Dev 15.

Ding, J., Yang, L., Yan, Y.T., Chen, A., Desai, N., Wynshaw-Boris, A., Shen, M.M., 1998. Cripto is required for correct orientation of the anterior-posterior axis in the mouse embryo. Nature 395, 702-707.

Dirksen, M.L., Jamrich, M., 1995. Differential expression of fork head genes during early Xenopus and zebrafish development. Developmental genetics 17, 107-116.

Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., 2013a. STAR: ultrafast universal RNA-seq aligner. Bioinformatics 29.

Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., Gingeras, T.R., 2013b. STAR: ultrafast universal RNA-seq aligner. Bioinformatics 29, 15-21.

Dobin, A., Gingeras, T.R., 2016. Optimizing RNA-Seq Mapping with STAR. Methods Mol Biol 1415, 245-262.

Dogan, A., 2018. Embryonic Stem Cells in Development and Regenerative Medicine. Advances in experimental medicine and biology 1079, 1-15.

Donà, E., Barry, J.D., Valentin, G., Quirin, C., Khmelinskii, A., Kunze, A., Durdu, S., Newton, L.R., Fernandez-Minan, A., Huber, W., Knop, M., Gilmour, D., 2013. Directional tissue migration through a self-generated chemokine gradient. Nature 503, 285.

Dorr, K.M., Amin, N.M., Kuchenbrod, L.M., Labiner, H., Charpentier, M.S., Pevny, L.H., Wessels, A., Conlon, F.L., 2015. Casz1 is required for cardiomyocyte G1-to-S phase progression during mammalian cardiac development. Development 142, 2037.

Dougan, S.T., Warga, R.M., Kane, D.A., Schier, A.F., Talbot, W.S., 2003. The role of the zebrafish nodal-related genes squint and cyclops in patterning of mesendoderm. Development 130, 1837-1851.

Driever, W., Stemple, D., Schier, A., Solnica-Krezel, L., 1994. Zebrafish: genetic tools for studying vertebrate development. Trends in Genetics 10, 152-159.

Du, S., Draper, B.W., Mione, M., Moens, C.B., Bruce, A., 2012. Differential regulation of epiboly initiation and progression by zebrafish Eomesodermin A. Dev Biol 362.

Dube, D.K., Dube, S., Abbott, L., Wang, J., Fan, Y., Alshiekh-Nasany, R., Shah, K.K., Rudloff, A.P., Poiesz, B.J., Sanger, J.M., Sanger, J.W., 2017. Identification, characterization, and expression of sarcomeric tropomyosin isoforms in zebrafish. Cytoskeleton 74, 125-142.

Dubrulle, J., Jordan, B.M., Akhmetova, L., Farrell, J.A., Kim, S.-H., Solnica-Krezel, L., Schier, A.F., 2015. Response to Nodal morphogen gradient is determined by the kinetics of target gene induction. eLife 4, e05042.

Dunham, I., Kundaje, A., Aldred, S.F., Collins, P.J., Davis, C.A., Doyle, F., Epstein, C.B., Frietze, S., Harrow, J., Kaul, R., Khatun, J., Lajoie, B.R., Landt, S.G., Lee, B.K., Pauli, F., Rosenbloom, K.R., Sabo, P., Safi, A., Shoresh, N., Simon, J.M., Song, L., Trinklein, N.D., Altshuler, R.C., Birney, E., Brown, J.B., Cheng, C., Djebali, S., Dong, X., 2012. An integrated encyclopedia of DNA elements in the human genome. Nature 489.

Dutton, K.A., Pauliny, A., Lopes, S.S., Elworthy, S., Carney, T.J., Rauch, J., Geisler, R., Haffter, P., Kelsh, R.N., 2001. Zebrafish colourless encodes sox10 and specifies non-ectomesenchymal neural crest fates. Development 128, 4113-4125.

Eijlander, R.T., Kuipers, O.P., 2013. Live-Cell Imaging Tool Optimization To Study Gene Expression Levels and Dynamics in Single Cells of Bacillus cereus. Applied and Environmental Microbiology 79, 5643.

El-Brolosy, M., Rossi, A., Kontarakis, Z., Kuenne, C., Guenther, S., Fukuda, N., Takacs, C., Lai, S.-L., Fukuda, R., Gerri, C., Kikhi, K., Giraldez, A., Stainier, D.Y.R., 2018. Genetic compensation is triggered by mutant mRNA degradation. bioRxiv, 328153.

El-Brolosy, M.A., Stainier, D.Y.R., 2017. Genetic compensation: A phenomenon in search of mechanisms. PLoS genetics 13, e1006780-e1006780.

Elnitski, L., Jin, V.X., Farnham, P.J., Jones, S.J., 2006. Locating mammalian transcription factor binding sites: a survey of computational and experimental techniques. Genome Res 16, 1455-1464.

Elsalini, O.A., von Gartzen, J., Cramer, M., Rohr, K.B., 2003. Zebrafish hhex, nk2.1a, and pax2.1 regulate thyroid growth and differentiation downstream of Nodal-dependent transcription factors. Dev Biol 263, 67-80.

Emmert-Streib, F., Dehmer, M., Haibe-Kains, B., 2014. Gene regulatory networks and their applications: understanding biological and medical problems in terms of networks. 2.

Engert, S., Burtscher, I., Liao, W.P., Dulev, S., Schotta, G., Lickert, H., 2013. Wnt/β-catenin signalling regulates Sox17 expression and is essential for organizer and endoderm formation in the mouse. Development 140, 3128.

Engleka, M.J., Craig, E.J., Kessler, D.S., 2001. VegT Activation of Sox17 at the Midblastula Transition Alters the Response to Nodal Signals in the Vegetal Endoderm Domain. Developmental Biology 237, 159-172.

Epstein, D.J., 2009. Cis-regulatory mutations in human disease. Briefings in functional genomics & proteomics 8, 310-316.

Erkenbrack, E.M., Davidson, E.H., Peter, I.S., 2018. Conserved regulatory state expression controlled by divergent developmental gene regulatory networks in echinoids. Development, dev.167288.

Erter, C.E., Solnica-Krezel, L., Wright, C.V., 1998. Zebrafish nodal-related 2 encodes an early mesendodermal inducer signaling from the extraembryonic yolk syncytial layer. Dev Biol 204.

Essner, J.J., Amack, J.D., Nyholm, M.K., Harris, E.B., Yost, H.J., 2005. Kupffer's vesicle is a ciliated organ of asymmetry in the zebrafish embryo that initiates left-right development of the brain, heart and gut. Development 132, 1247-1260.

Etienne-Manneville, S., 2004. Cdc42 - the centre of polarity. Journal of cell science 117, 1291.

Fan, X., Hagos, E.G., Xu, B., Sias, C., Kawakami, K., Burdine, R.D., Dougan, S.T., 2007. Nodal signals mediate interactions between the extra-embryonic and embryonic tissues in zebrafish. Developmental Biology 310, 363-378.

Farley, E.K., Olson, K.M., Zhang, W., Brandt, A.J., Rokhsar, D.S., Levine, M.S., 2015. Suboptimization of developmental enhancers. Science 350, 325-328.

Farrell, J.A., Wang, Y., Riesenfeld, S.J., Shekhar, K., Regev, A., Schier, A.F., 2018. Single-cell reconstruction of developmental trajectories during zebrafish embryogenesis. Science 360.

Feldman, B., Dougan, S.T., Schier, A.F., Talbot, W.S., 2000. Nodal-related signals establish mesendodermal fate and trunk neural identity in zebrafish. Current Biology 10, 531-534.

Feldman, B., Gates, M.A., Egan, E.S., Dougan, S.T., Rennebeck, G., Sirotkin, H.I., Schier, A.F., Talbot, W.S., 1998. Zebrafish organizer development and germ-layer formation require nodal-related signals. Nature 395, 181-185.

Feng, J., Liu, T., Zhang, Y., 2011. Using MACS to Identify Peaks from ChIP-Seq Data. Current protocols in bioinformatics / editoral board, Andreas D. Baxevanis ... [et al.] CHAPTER, Unit2.14-Unit12.14.

Fernandez-Minan, A., Bessa, J., Tena, J.J., Gomez-Skarmeta, J.L., 2016. Assay for transposase-accessible chromatin and circularized chromosome conformation capture, two methods to explore the regulatory landscapes of genes in zebrafish. Methods in cell biology 135, 413-430.

Ferrell, James E., Jr., 2012. Bistability, Bifurcations, and Waddington's Epigenetic Landscape. Current Biology 22, R458-R466.

Fisher, J.B., Pulakanti, K., Rao, S., Duncan, S.A., 2017. GATA6 is essential for endoderm formation from human pluripotent stem cells. Biology Open 6, 1084.

Flowers, G.P., Topczewska, J.M., Topczewski, J., 2012. A zebrafish Notum homolog specifically blocks the Wnt/beta-catenin signaling pathway. Development 139.

FlyBase, C., Thurmond, J., Goodman, J.L., Kaufman, T.C., Strelets, V.B., Calvi, B.R., Millburn, G., Antonazzo, G., Attrill, H., Marygold, S.J., Trovisco, V., Matthews, B.B., Gramates, L S., 2018. FlyBase 2.0: the next generation. Nucleic Acids Research 47, D759-D765.

Foulk, M.S., Urban, J.M., Casella, C., Gerbi, S.A., 2015. Characterizing and controlling intrinsic biases of lambda exonuclease in nascent strand sequencing reveals phasing between nucleosomes and G-quadruplex motifs around a subset of human replication origins. Genome research 25, 725-735.

Francois, M., Koopman, P., Beltrame, M., 2010. SoxF genes: Key players in the development of the cardio-vascular system. Int J Biochem Cell Biol 42, 445-448.

Fuerer, C., Nostro, M.C., Constam, D.B., 2014. Nodal Gdf1 heterodimers with bound prodomains enable serum-independent nodal signaling and endoderm differentiation. J Biol Chem 289, 17854-17871.

Furey, T.S., 2012. ChIP–seq and beyond: new and improved methodologies to detect and characterize protein–DNA interactions. Nature Reviews Genetics 13, 840.

Gaither, M.R., Gkafas, G.A., de Jong, M., Sarigol, F., Neat, F., Regnier, T., Moore, D., Gröcke, D.R., Hall, N., Liu, X., Kenny, J., Lucaci, A., Hughes, M., Haldenby, S., Hoelzel, A.R., 2018. Genomics of habitat choice and adaptive evolution in a deep-sea fish. Nature Ecology & Evolution 2, 680-687.

Gallardo, V.E., Behra, M., 2013. Fluorescent activated cell sorting (FACS) combined with gene expression microarrays for transcription enrichment profiling of zebrafish lateral line cells. Methods 62, 226-231.

Gallego Romero, I., Pai, A.A., Tung, J., Gilad, Y., 2014. RNA-seq: impact of RNA degradation on transcript quantification. BMC biology 12, 42-42.

Gao, C., Lo, L.J., Huang, W., Gao, Y., Peng, J., Huang, H., Luo, L., Chen, J., 2018. Zebrafish hhex-null mutant develops an intrahepatic intestinal tube due to de-repression of cdx1b and pdx1.

Gao, N., LeLay, J., Vatamaniuk, M.Z., Rieck, S., Friedman, J.R., Kaestner, K.H., 2008. Dynamic regulation of Pdx1 enhancers by Foxa1 and Foxa2 is essential for pancreas development. 22, 3435-3448.

Garcia-Elias, A., Alloza, L., Puigdecanet, E., Nonell, L., Tajes, M., Curado, J., Enjuanes, C., Díaz, O., Bruguera, J., Martí-Almor, J., Comín-Colet, J., Benito, B., 2017. Defining quantification methods and optimizing protocols for microarray hybridization of circulating microRNAs. Scientific reports 7, 7725-7725.

Garcia-Fernàndez, J., Pascual-Anaya, J., Jiménez-Delgado, S., 2009. Implications of duplicated cis-regulatory elements in the evolution of metazoans: the DDI model or how simplicity begets novelty. Briefings in Functional Genomics 8, 266-275.

Gehrig, J., Reischl, M., Kalmar, E., Ferg, M., Hadzhiev, Y., Zaucker, A., Song, C., Schindler, S., Liebel, U., Muller, F., 2009. Automated high-throughput mapping of promoter-enhancer interactions in zebrafish embryos. Nat Methods 6, 911-916.

Gentsch, G.E., Owens, N.D., Martin, S.R., Piccinelli, P., Faial, T., Trotter, M.W., Gilchrist, M.J., Smith, J.C., 2013. In vivo T-box transcription factor profiling reveals joint regulation of embryonic neuromesodermal bipotency. Cell Rep 4.

Gentsch, G.E., Patrushev, I., Smith, J.C., 2015. Genome-wide snapshot of chromatin regulators and states in Xenopus embryos by ChIP-Seq. J Vis Exp.

Gentsch, G.E., Smith, J.C., 2019. Mapping Chromatin Features of Xenopus Embryos. Cold Spring Harb Protoc 2019, pdb.prot100263.

Gentsch, G.E., Spruce, T., Monteiro, R.S., Owens, N.D.L., Martin, S.R., Smith, J.C., 2018. Innate Immune Response and Off-Target Mis-splicing Are Common Morpholino-Induced Side Effects in Xenopus. Dev Cell 44, 597-610 e510.

Gerke, J., Lorenz, K., Cohen, B., 2009. Genetic interactions between transcription factors cause natural variation in yeast. Science 323, 498-501.

Gerstein, M.B., Lu, Z.J., Van Nostrand, E.L., Cheng, C., Arshinoff, B.I., Liu, T., Yip, K.Y., Robilotto, R., Rechtsteiner, A., Ikegami, K., Alves, P., Chateigner, A., Perry, M., Morris, M., Auerbach, R.K., Feng, X., Leng, J., Vielle, A., Niu, W., Rhrissorrakrai, K., Agarwal, A., Alexander, R.P., Barber, G., Brdlik, C.M., Brennan, J., Brouillet, J.J., Carr, A., Cheung, M.S., Clawson, H., Contrino, S., Dannenberg, L.O., Dernburg, A.F., Desai, A., Dick, L., Dose, A.C., Du, J., Egelhofer, T., Ercan, S., Euskirchen, G., Ewing, B., Feingold, E.A., Gassmann, R., Good, P.J., Green, P., Gullier, F., Gutwein, M., Guyer, M.S., Habegger, L., Han, T., Henikoff, J.G., Henz, S.R., Hinrichs, A., Holster, H., Hyman, T., Iniguez, A.L., Janette, J., Jensen, M., Kato, M., Kent, W.J., Kephart, E., Khivansara, V., Khurana, E., Kim, J.K., Kolasinska-Zwierz, P., Lai, E.C., Latorre, I., Leahey, A., Lewis, S., Lloyd, P., Lochovsky, L., Lowdon, R.F., Lubling, Y., Lyne, R., MacCoss, M., Mackowiak, S.D., Mangone, M., McKay, S., Mecenas, D., Merrihew, G., Miller, D.M., 3rd, Muroyama, A., Murray, J.I., Ooi, S.L., Pham, H., Phippen, T., Preston, E.A., Rajewsky, N., Ratsch, G., Rosenbaum, H., Rozowsky, J., Rutherford, K., Ruzanov, P., Sarov, M., Sasidharan, R., Sboner, A., Scheid, P., Segal, E., Shin, H., Shou, C., Slack, F.J., Slightam, C., Smith, R., Spencer, W.C., Stinson, E.O., Taing, S., Takasaki, T., Vafeados, D., Voronina, K., Wang, G., Washington, N.L., Whittle, C.M., Wu, B., Yan, K.K., Zeller, G., Zha, Z., Zhong, M., Zhou, X., Ahringer, J., Strome, S., Gunsalus, K.C., Micklem,

G., Liu, X.S., Reinke, V., Kim, S.K., Hillier, L.W., Henikoff, S., Piano, F., Snyder, M., Stein, L., Lieb, J.D., Waterston, R.H., 2010. Integrative analysis of the Caenorhabditis elegans genome by the modENCODE project. Science 330, 1775-1787.

Giger, F.A., David, N.B., 2017. Endodermal germ-layer formation through active actin-driven migration triggered by N-cadherin. Proceedings of the National Academy of Sciences 114, 10143.

Gligorijević, V., Pržulj, N., 2015. Methods for biological data integration: perspectives and challenges. Journal of the Royal Society, Interface 12, 20150571.

Golson, M.L., Kaestner, K.H., 2016. Fox transcription factors: from development to disease. Development (Cambridge, England) 143, 4558-4570.

Gong, Z., Ju, B., Wan, H., 2001. Green fluorescent protein (GFP) transgenic fish and their applications. Genetica 111, 213-225.

Grapin-Botton, A., Constam, D., 2007. Evolution of the mechanisms and molecular control of endoderm formation. Mechanisms of development 124, 253-278.

Greenhill, E.R., Rocco, A., Vibert, L., Nikaido, M., Kelsh, R.N., 2011. An Iterative Genetic and Dynamical Modelling Approach Identifies Novel Features of the Gene Regulatory Network Underlying Melanocyte Development. PLOS Genetics 7, e1002265.

Gritsman, K., Talbot, W.S., Schier, A.F., 2000. Nodal signaling patterns the organizer. Development 127, 921-932.

Gritsman, K., Zhang, J., Cheng, S., Heckscher, E., Talbot, W.S., Schier, A.F., 1999. The EGF-CFC protein one-eyed pinhead is essential for nodal signaling. Cell 97.

Gronenborn, A.M., 2005. The DNA-Binding Domain of GATA Transcription Factors—A Prototypical Type IV Cys2-Cys2 Zinc Finger, in: Iuchi, S., Kuldell, N. (Eds.), Zinc Finger Proteins: From Atomic Contact to Cellular Function. Springer US, Boston, MA, pp. 26-30.

Guo, Y., Mahony, S., Gifford, D.K., 2012. High Resolution Genome Wide Binding Event Finding and Motif Discovery Reveals Transcription Factor Spatial Binding Constraints. PLoS Comput Biol 8, e1002638.

Gurgul, A., Pawlina-Tyszko, K., Bugno-Poniewierska, M., Szmato, #x, a, T., Jasielczuk, I., Dobosz, S., Ocalewicz, K., 2018. Transcriptome Analysis of Rainbow Trout (Oncorhynchus mykiss) Eggs Subjected to the High Hydrostatic Pressure Treatment %J International Journal of Genomics. 2018, 7.

Hall, C., Flores, M.V., Murison, G., Crosier, K., Crosier, P., 2006. An essential role for zebrafish Fgfrl1 during gill cartilage development. Mechanisms of development 123, 925-940.

Hardcastle, T.J., Kelly, K.A., 2010. baySeq: empirical Bayesian methods for identifying differential expression in sequence count data. BMC Bioinformatics 11.

Haring, M., Offermann, S., Danker, T., Horst, I., Peterhansel, C., Stam, M., 2007. Chromatin immunoprecipitation: optimization, quantitative analysis and data normalization. Plant methods 3, 11-11.

Harley, V.R., Lovell-Badge, R., Goodfellow, P.N., 1994. Definition of a consensus DNA binding site for SRY. Nucleic Acids Res 22, 1500-1501.

Hartonen, T., Sahu, B., Dave, K., Kivioja, T., Taipale, J., 2016. PeakXus: comprehensive transcription factor binding site discovery from ChIP-Nexus and ChIP-Exo experiments. Bioinformatics 32, i629-i638.

Harvey, S.A., Sealy, I., Kettleborough, R., Fenyes, F., White, R., Stemple, D., Smith, J.C., 2013. Identification of the zebrafish maternal and paternal transcriptomes. Development 140, 2703.

Hashimshony, T., Feder, M., Levin, M., Hall, B.K., Yanai, I., 2015. Spatiotemporal transcriptomics reveals the evolutionary history of the endoderm germ layer. Nature 519, 219-222.

Hawkins, J.R., 1993. Mutational analysis of SRY in XY females. Human mutation 2, 347-350.

He, A., Pu, W.T., 2010. Genome-wide location analysis by pull down of in vivo biotinylated transcription factors. Current protocols in molecular biology Chapter 21, Unit-21.20.

He, L., Zhang, A., Pei, Y., Chu, P., Li, Y., Huang, R., Liao, L., Zhu, Z., Wang, Y., 2017. Differences in responses of grass carp to different types of grass carp reovirus (GCRV) and the mechanism of hemorrhage revealed by transcriptome sequencing. BMC genomics 18, 452-452.

He, Q., Johnston, J., Zeitlinger, J., 2015. ChIP-nexus enables improved detection of in vivo transcription factor binding footprints. Nat Biotechnol 33, 395-401.

He, X., Zhang, J., 2005. Rapid subfunctionalization accompanied by prolonged and substantial neofunctionalization in duplicate gene evolution. Genetics 169, 1157-1164.

Heenan, P., Zondag, L., Wilson, M.J., 2016. Evolution of the Sox gene family within the chordate phylum. Gene 575, 385-392.

Heicklen-Klein, A., McReynolds, L.J., Evans, T., 2005. Using the zebrafish model to study GATA transcription factors. Seminars in Cell & Developmental Biology 16, 95-106.

Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y.C., Laslo, P., Cheng, J.X., Murre, C., Singh, H., Glass, C.K., 2010. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. Mol Cell 38, 576-589.

Hellman, N.E., Liu, Y., Merkel, E., Austin, C., Le Corre, S., Beier, D.R., Sun, Z., Sharma, N., Yoder, B.K., Drummond, I.A., 2010. The zebrafish foxj1a transcription factor regulates cilia function in response to injury and epithelial stretch. Proceedings of the National Academy of Sciences 107, 18499.

Hermsen, R., Ursem, B., ten Wolde, P.R., 2010. Combinatorial Gene Regulation Using Auto-Regulation. PLOS Computational Biology 6, e1000813.

Herpers, R., van de Kamp, E., Duckers, H.J., Schulte-Merker, S., 2008. Redundant Roles for Sox7 and Sox18 in Arteriovenous Specification in Zebrafish. Circulation Research 102, 12-15.

Herpin, A., Schartl, M., 2015. Plasticity of gene-regulatory networks controlling sex determination: of masters, slaves, usual suspects, newcomers, and usurpators. EMBO reports 16, 1260.

Higashijima, S., 2008. Transgenic zebrafish expressing fluorescent proteins in central nervous system neurons. Dev Growth Differ 50, 407-413.

Hill, C.S., 2018. Spatial and temporal control of NODAL signaling. Current opinion in cell biology 51, 50-57.

Hinits, Y., Pan, L., Walker, C., Dowd, J., Moens, C.B., Hughes, S.M., 2012. Zebrafish Mef2ca and Mef2cb are essential for both first and second heart field cardiomyocyte differentiation. Dev Biol 369, 199-210.

Hinton, A., Afrikanova, I., Wilson, M., King, C.C., Maurer, B., Yeo, G.W., Hayek, A., Pasquinelli, A.E., 2010. A distinct microRNA signature for definitive endoderm derived from human embryonic stem cells. Stem cells and development 19, 797-807.

Hirata, T., Yamanaka, Y., Ryu, S.-L., Shimizu, T., Yabe, T., Hibi, M., Hirano, T., 2000. Novel Mix-Family Homeobox Genes in Zebrafish and Their Differential Regulation. Biochemical and biophysical research communications 271, 603-609.

Ho, R.K., Kimmel, C.B., 1993. Commitment of cell fate in the early zebrafish embryo. Science 261, 109-111.

Hoffman, B.G., Jones, S.J., 2009. Genome-wide identification of DNA-protein interactions using chromatin immunoprecipitation coupled with flow cell sequencing. J Endocrinol 201, 1-13.

Holtzinger, A., Evans, T., 2005. Gata4 regulates the formation of multiple organs. Development 132, 4005.

Holtzinger, A., Evans, T., 2007. Gata5 and Gata6 are functionally redundant in zebrafish for specification of cardiomyocytes. Developmental biology 312, 613-622.

Holtzinger, A., Rosenfeld, G.E., Evans, T., 2010. Gata4 directs development of cardiac-inducing endoderm from ES cells. Developmental biology 337, 63-73.

Hosking, B.M., Wang, S.C., Chen, S.L., Penning, S., Koopman, P., Muscat, G.E., 2001. SOX18 directly interacts with MEF2C in endothelial cells. Biochemical and biophysical research communications 287, 493-500.

Hostelley, T.L., Nesmith, J.E., Zaghloul, N.A., 2017. Sample Preparation and Analysis of RNASeq-based Gene Expression Data from Zebrafish. J Vis Exp.

Howe, K., Clark, M.D., Torroja, C.F., Torrance, J., Berthelot, C., Muffato, M., Collins, J.E., Humphray, S., McLaren, K., Matthews, L., McLaren, S., Sealy, I., Caccamo, M., Churcher, C., Scott, C., Barrett, J.C., Koch, R., Rauch, G.-J., White, S., Chow, W., Kilian, B., Quintais, L.T., Guerra-Assunção, J.A., Zhou, Y., Gu, Y., Yen, J., Vogel, J.-H., Eyre, T., Redmond, S., Banerjee, R., Chi, J., Fu, B., Langley, E., Maguire, S.F., Laird, G.K., Lloyd, D., Kenyon, E., Donaldson, S., Sehra, H., Almeida-King, J., Loveland, J., Trevanion, S., Jones, M., Quail, M., Willey, D., Hunt, A., Burton, J., Sims, S., McLay, K., Plumb, B., Davis, J., Clee, C., Oliver,

K., Clark, R., Riddle, C., Elliott, D., Threadgold, G., Harden, G., Ware, D., Begum, S., Mortimore, B., Kerry, G., Heath, P., Phillimore, B., Tracey, A., Corby, N., Dunn, M., Johnson, C., Wood, J., Clark, S., Pelan, S., Griffiths, G., Smith, M., Glithero, R., Howden, P., Barker, N., Lloyd, C., Stevens, C., Harley, J., Holt, K., Panagiotidis, G., Lovell, J., Beasley, H., Henderson, C., Gordon, D., Auger, K., Wright, D., Collins, J., Raisen, C., Dyer, L., Leung, K., Robertson, L., Ambridge, K., Leongamornlert, D., McGuire, S., Gilderthorp, R., Griffiths, C., Manthravadi, D., Nichol, S., Barker, G., Whitehead, S., Kay, M., Brown, J., Murnane, C., Gray, E., Humphries, M., Sycamore, N., Barker, D., Saunders, D., Wallis, J., Babbage, A., Hammond, S., Mashreghi-Mohammadi, M., Barr, L., Martin, S., Wray, P., Ellington, A., Matthews, N., Ellwood, M., Woodmansey, R., Clark, G., Cooper, J.D., Tromans, A., Grafham, D., Skuce, C., Pandian, R., Andrews, R., Harrison, E., Kimberley, A., Garnett, J., Fosker, N., Hall, R., Garner, P., Kelly, D., Bird, C., Palmer, S., Gehring, I., Berger, A., Dooley, C.M., Ersan-Ürün, Z., Eser, C., Geiger, H., Geisler, M., Karotki, L., Kirn, A., Konantz, J., Konantz, M., Oberländer, M., Rudolph-Geiger, S., Teucke, M., Lanz, C., Raddatz, G., Osoegawa, K., Zhu, B., Rapp, A., Widaa, S., Langford, C., Yang, F., Schuster, S.C., Carter, N.P., Harrow, J., Ning, Z., Herrero, J., Searle, S.M.J., Enright, A., Geisler, R., Plasterk, R.H.A., Lee, C., Westerfield, M., de Jong, P.J., Zon, L.I., Postlethwait, J.H., Nüsslein-Volhard, C., Hubbard, T.J.P., Crollius, H.R., Rogers, J., Stemple, D.L., 2013. The zebrafish reference genome sequence and its relationship to the human genome. Nature 496, 498.

Howell, M., Mohun, T.J., Hill, C.S., 2001. Xenopus Smad3 is specifically expressed in the chordoneural hinge, notochord and in the endocardium of the developing heart. Mechanisms of development 104, 147-150.

Huang da, W., Sherman, B.T., Lempicki, R.A., 2009. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc 4, 44-57.

Hudson, C., Clements, D., Friday, R.V., Stott, D., Woodland, H.R., 1997. Xsox17alpha and -beta mediate endoderm formation in Xenopus. Cell 91, 397-405.

Hug, C.B., Grimaldi, A.G., Kruse, K., Vaquerizas, J.M., 2017. Chromatin Architecture Emerges during Zygotic Genome Activation Independent of Transcription. Cell 169, 216-228.e219.

Huttenlocher, A., Horwitz, A.R., Integrins in cell migration. Cold Spring Harbor perspectives in biology 3, a005074-a005074.

Illumina, 2011. RNA-Seq Data Comparison with Gene Expression Microarrays. http://www.europeanpharmaceuticalreview.com/wp-content/uploads/Illumina_whitepaper.pdf.

Imai, K.S., Stolfi, A., Levine, M., Satou, Y., 2009. Gene regulatory networks underlying the compartmentalization of the Ciona central nervous system. Development 136, 285.

Imai, Y., Gates, M.A., Melby, A.E., Kimelman, D., Schier, A.F., Talbot, W.S., 2001. The homeobox genes vox and vent are redundant repressors of dorsal fates in zebrafish. Development 128, 2407-2420.

Inoue, T., Inoue, Y.U., Asami, J., Izumi, H., Nakamura, S., Krumlauf, R., 2008. Analysis of mouse Cdh6 gene regulation by transgenesis of modified bacterial artificial chromosomes. Developmental Biology 315, 506-520.

Ishikawa, D., Diekmann, U., Fiedler, J., Just, A., Thum, T., Lenzen, S., Naujok, O., 2017. miRNome Profiling of Purified Endoderm and Mesoderm Differentiated from hESCs Reveals Functions of miR-483-3p and miR-1263 for Cell-Fate Decisions. Stem cell reports 9, 1588-1603.

Jahangiri, L., Nelson, A.C., Wardle, F.C., 2012. A cis-regulatory module upstream of deltaC regulated by Ntla and Tbx16 drives expression in the tailbud, presomitic mesoderm and somites. Developmental biology 371, 110-120.

James-Zorn, C., Ponferrada, V.G., Burns, K.A., Fortriede, J.D., Lotay, V.S., Liu, Y., Brad Karpinka, J., Karimi, K., Zorn, A.M., Vize, P.D., 2015. Xenbase: Core features, data acquisition, and data processing. Genesis 53, 486-497.

Ji, Y., Chae, S., Lee, H.-K., Park, I., Kim, C., Ismail, T., Kim, Y., Park, J.-W., Kwon, O.-S., Kang, B.-S., Lee, D.-S., Bae, J.-S., Kim, S.-H., Moon, P.-G., Baek, M.-C., Park, M.-J., Kil, I.S., Rhee, S.G., Kim, J., Huh, Y.H., Shin, J.-Y., Min, K.-J., Kwon, T.K., Jang, D.G., Woo, H.A., Kwon, T., Park, T.J., Lee, H.-S., 2018. Peroxiredoxin5 Controls Vertebrate Ciliogenesis by Modulating Mitochondrial Reactive Oxygen Species. Antioxidants & Redox Signaling.

Johnson, D.S., Mortazavi, A., Myers, R.M., Wold, B., 2007. Genome-Wide Mapping of in Vivo Protein-DNA Interactions. Science 316, 1497-1502.

Jones, B.L., Swallow, D.M., 2011. The impact of cis-acting polymorphisms on the human phenotype. The HUGO journal 5, 13-23.

Junker, Jan P., Noël, Emily S., Guryev, V., Peterson, Kevin A., Shah, G., Huisken, J., McMahon, Andrew P., Berezikov, E., Bakkers, J., van Oudenaarden, A., 2014. Genome-wide RNA Tomography in the Zebrafish Embryo. Cell 159, 662-675.

Kaaij, L.J.T., Mokry, M., Zhou, M., Musheev, M., Geeven, G., Melquiond, A.S.J., de Jesus Domingues, A.M., de Laat, W., Niehrs, C., Smith, A.D., Ketting, R.F., 2016. Enhancers reside in a unique epigenetic environment during early zebrafish development. Genome Biology 17, 146.

Kaaij, L.J.T., van der Weide, R.H., Ketting, R.F., de Wit, E., 2018. Systemic Loss and Gain of Chromatin Architecture throughout Zebrafish Development. Cell Rep 24, 1-10 e14.

Kałużna, M., Kuras, A., Mikiciński, A., Puławska, J., 2016. Evaluation of different RNA extraction methods for high-quality total RNA and mRNA from Erwinia amylovora in planta. European Journal of Plant Pathology 146, 893-899.

Kamachi, Y., Kondoh, H., 2013. Sox proteins: regulators of cell fate specification and differentiation. Development 140, 4129-4144.

Kamachi, Y., Uchikawa, M., Tanouchi, A., Sekido, R., Kondoh, H., 2001. Pax6 and SOX2 form a co-DNA-binding partner complex that regulates initiation of lens development. Genes & Development 15, 1272-1286.

Kanai-Azuma, M., Kanai, Y., Gad, J.M., Tajima, Y., Taya, C., Kurohmaru, M., Sanai, Y., Yonekawa, H., Yazaki, K., Tam, P.P., Hayashi, Y., 2002. Depletion of definitive gut endoderm in Sox17-null mutant mice. Development 129, 2367-2379.

Kassahn, K.S., Dang, V.T., Wilkins, S.J., Perkins, A.C., Ragan, M.A., 2009. Evolution of gene function and regulatory control after whole-genome duplication: comparative analyses in vertebrates. Genome research 19, 1404-1418.

Keller, R., 2005. Cell migration during gastrulation. Current opinion in cell biology 17, 533-541.

Kidder, B.L., Hu, G., Zhao, K., 2011. ChIP-Seq: technical considerations for obtaining high-quality data. Nature immunology 12, 918-922.

Kiecker, C., Bates, T., Bell, E., 2016. Molecular specification of germ layers in vertebrate embryos. Cell Mol Life Sci 73, 923-947.

Kikuchi, K., Holdway, Jennifer E., Major, Robert J., Blum, N., Dahn, Randall D., Begemann, G., Poss, Kenneth D., 2011. Retinoic Acid Production by Endocardium and Epicardium Is an Injury Response Essential for Zebrafish Heart Regeneration. Developmental Cell 20, 397-404.

Kikuchi, Y., Agathon, A., Alexander, J., Thisse, C., Waldron, S., Yelon, D., Thisse, B., Stainier, D.Y., 2001. Casanova encodes a novel Sox-related protein necessary and sufficient for early endoderm formation in zebrafish. Genes Dev 15.

Kikuchi, Y., Trinh, L.A., Reiter, J.F., Alexander, J., Yelon, D., Stainier, D.Y.R., 2000. The zebrafish bonnie and clyde gene encodes a Mix family homeodomain protein that regulates the generation of endodermal precursors. Genes & Development 14, 1279-1289.

Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R., Salzberg, S., 2013. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol 14.

Kimelman, D., Griffin, K.J.P., 2000. Vertebrate mesendoderm induction and patterning. Current Opinion in Genetics & Development 10, 350-356.

Kimmel, C.B., Ballard, W.W., Kimmel, S.R., Ullmann, B., Schilling, T.F., 1995. Stages of embryonic development of the zebrafish. Developmental Dynamics 203, 253-310.

Kimmel, C.B., Warga, R.M., Schilling, T.F., 1990. Origin and organization of the zebrafish fate map. Development 108, 581-594.

Kinkel, M.D., Eames, S.C., Alonzo, M.R., Prince, V.E., 2008. Cdx4 is required in the endoderm to localize the pancreas and limit beta-cell number. Development 135, 919-929.

Kleinjan, D.A., Bancewicz, R.M., Gautier, P., Dahm, R., Schonthaler, H.B., Damante, G., Seawright, A., Hever, A.M., Yeyati, P.L., van Heyningen, V., Coutinho, P., 2008. Subfunctionalization of Duplicated Zebrafish pax6 Genes by cis-Regulatory Divergence. PLOS Genetics 4, e29.

Klüver, N., Kondo, M., Herpin, A., Mitani, H., Schartl, M., 2005. Divergent expression patterns of Sox9 duplicates in teleosts indicate a lineage specific subfunctionalization. Development genes and evolution 215, 297-305.

Knight, R.D., Nair, S., Nelson, S.S., Afshar, A., Javidan, Y., Geisler, R., Rauch, G.J., Schilling, T.F., 2003. Lockjaw encodes a zebrafish tfap2a required for early neural crest development. Development 130.

Kobayashi, D., Jindo, T., Naruse, K., Takeda, H., 2006. Development of the endoderm and gut in medaka, Oryzias latipes. Dev Growth Differ 48, 283-295.

Koopman, P., Gubbay, J., Vivian, N., Goodfellow, P., Lovell-Badge, R., 1991. Male development of chromosomally female mice transgenic for Sry. Nature 351, 117-121.

Kopylova, E., Noe, L., Touzet, H., 2012. SortMeRNA: fast and accurate filtering of ribosomal RNAs in metatranscriptomic data. Bioinformatics 28, 3211-3217.

Koshida, S., Shinya, M., Mizuno, T., Kuroiwa, A., Takeda, H., 1998. Initial anteroposterior pattern of the zebrafish central nervous system is determined by differential competence of the epiblast. Development 125, 1957.

Koster, R., Sassen, W., 2015. A molecular toolbox for genetic manipulation of zebrafish. Advances in Genomics and Genetics 5, 151.

Krueger, F., 2012. Trim Galore! . Barbraham Bioinformatics.

Kruse, K., Diaz, N., Enriquez-Gasca, R., Gaume, X., Torres-Padilla, M.-E., Vaquerizas, J.M., 2019. Transposable elements drive reorganisation of 3D chromatin during early embryogenesis. bioRxiv, 523712.

Kubo, A., Suzuki, N., Yuan, X., Nakai, K., Satoh, N., Imai, K.S., Satou, Y., 2010. Genomic cis-regulatory networks in the early Ciona intestinalis embryo. Development 137, 1613.

Kudtarkar, P., Cameron, R.A., 2017. Echinobase: an expanding resource for echinoderm genomic information. Database : the journal of biological databases and curation 2017, bax074.

Kukurba, K.R., Montgomery, S.B., 2015. RNA Sequencing and Analysis. Cold Spring Harbor protocols 2015, 951-969.

Lamarre, S., Frasse, P., Zouine, M., Labourdette, D., Sainderichin, E., Hu, G., Le Berre-Anton, V., Bouzayen, M., Maza, E., 2018. Optimization of an RNA-Seq Differential Gene Expression Analysis Depending on Biological Replicate Number and Library Size. 9.

LaMonica, K., Ding, H.L., Artinger, K.B., 2015. prdm1a functions upstream of itga5 in zebrafish craniofacial development. Genesis 53, 270-277.

Lan, X., Pritchard, J.K., 2016. Coregulation of tandem duplicate genes slows evolution of subfunctionalization in mammals. Science 352, 1009-1013.

Landt, S.G., Marinov, G.K., Kundaje, A., Kheradpour, P., Pauli, F., Batzoglou, S., Bernstein, B.E., Bickel, P., Brown, J.B., Cayting, P., Chen, Y., DeSalvo, G., Epstein, C., Fisher-Aylor, K.I., Euskirchen, G., Gerstein, M., Gertz, J., Hartemink, A.J., Hoffman, M.M., Iyer, V.R., Jung, Y.L., Karmakar, S., Kellis, M., Kharchenko, P.V., Li, Q., Liu, T., Liu, X.S., Ma, L., Milosavljevic, A., Myers, R.M., Park, P.J., Pazin, M.J., Perry, M.D., Raha, D., Reddy, T.E., Rozowsky, J., Shoresh, N., Sidow, A., Slattery, M., Stamatoyannopoulos, J.A., Tolstorukov, M.Y., White, K.P., Xi, S., Farnham, P.J., Lieb, J.D., Wold, B.J., Snyder, M., 2012. ChIP-seq

guidelines and practices of the ENCODE and modENCODE consortia. Genome Research 22, 1813-1831.

Langmead, B., 2010. Aligning short sequencing reads with Bowtie. Curr Protoc Bioinformatics Chapter 11, Unit 11 17.

Langmead, B., Salzberg, S.L., 2012. Fast gapped-read alignment with Bowtie 2. Nat Meth 9, 357-359.

Larionov, A., Krause, A., Miller, W., 2005. A standard curve based method for relative real time PCR data processing. BMC bioinformatics 6, 62-62.

Latimer, A.J., Jessen, J.R., 2008. hgf/c-met expression and functional analysis during zebrafish embryogenesis. 237, 3904-3915.

Lawrence, M., Huber, W., Pages, H., Aboyoun, P., Carlson, M., Gentleman, R., Morgan, M.T., Carey, V.J., 2013. Software for computing and annotating genomic ranges. PLoS Comput Biol 9, e1003118.

Lazic, S., Scott, I.C., 2011. Mef2cb regulates late myocardial cell addition from a second heart field-like population of progenitors in zebrafish. Dev Biol 354, 123-133.

Lee, C.S., Friedman, J.R., Fulmer, J.T., Kaestner, K.H., 2005. The initiation of liver development is dependent on Foxa transcription factors. Nature 435, 944-947.

Lee, M.T., Bonneau, A.R., Takacs, C.M., Bazzini, A.A., DiVito, K.R., Fleming, E.S., Giraldez, A.J., 2013. Nanog, Pou5f1 and SoxB1 activate zygotic gene expression during the maternal-to-zygotic transition. Nature 503, 360-364.

Lee, T.I., Rinaldi, N.J., Robert, F., Odom, D.T., Bar-Joseph, Z., Gerber, G.K., Hannett, N.M., Harbison, C.T., Thompson, C.M., Simon, I., 2002. Transcriptional regulatory networks in Saccharomyces cerevisiae. Science 298.

Leek JT, J.W., Parker HS, Fertig EJ, Jaffe AE, Storey JD, Zhang Y, Torres LC, 2019. sva: Surrogate Variable Analysis. . R package version 3.30.1.

Leichsenring, M., Maes, J., Mossner, R., Driever, W., Onichtchouk, D., 2013. Pou5f1 transcription factor controls zygotic gene activation in vertebrates. Science 341, 1005-1009.

Leng, N., Dawson, J.A., Thomson, J.A., Ruotti, V., Rissman, A.I., Smits, B.M., 2013. EBSeq: an empirical Bayes hierarchical model for inference in RNA-seq experiments. Bioinformatics 29.

Levine, M., Davidson, E.H., 2005. Gene regulatory networks for development. Proc Natl Acad Sci USA 102.

Lewis, S.L., Tam, P.P.L., 2006. Definitive endoderm of the mouse embryo: Formation, cell fates, and morphogenetic function. Developmental Dynamics 235, 2315-2329.

Li, E., Davidson, E.H., 2009. Building developmental gene regulatory networks. Birth defects research. Part C, Embryo today : reviews 87, 123-130.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25, 2078-2079.

Li, N., Wei, C., Olena, A.F., Patton, J.G., 2011a. Regulation of endoderm formation and left-right asymmetry by miR-92 during early zebrafish development. Development (Cambridge, England) 138, 1817-1826.

Li, Q., Brown, J.B., Huang, H., Bickel, P.J., 2011b. Measuring reproducibility of high-throughput experiments. The Annals of Applied Statistics 5, 1752-1779.

Li, W., Cornell, R.A., 2007. Redundant activities of Tfap2a and Tfap2c are required for neural crest induction and development of other non-neural ectoderm derivatives in zebrafish embryos. Dev Biol 304.

Li, W.V., Li, J.J., 2018. Modeling and analysis of RNA-seq data: a review from a statistical perspective. Quantitative Biology 6, 195-209.

Liao, Y., Smyth, G.K., Shi, W., 2014. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics 30.

Lim, H.W., Uhlenhaut, N.H., Rauch, A., Weiner, J., Hubner, S., Hubner, N., Won, K.J., Lazar, M.A., Tuckermann, J., Steger, D.J., 2015. Genomic redistribution of GR monomers and dimers mediates transcriptional response to exogenous glucocorticoid in vivo. Genome Res 25, 836-844.

Lim, S.M., Pereira, L., Wong, M.S., Hirst, C.E., Van Vranken, B.E., Pick, M., Trounson, A., Elefanty, A.G., Stanley, E.G., 2009. Enforced expression of Mixl1 during mouse ES cell differentiation suppresses hematopoietic mesoderm and promotes endoderm formation. Stem cells (Dayton, Ohio) 27, 363-374.

Linde, J., Schulze, S., Henkel, S.G., Guthke, R., 2015. Data- and knowledge-based modeling of gene regulatory networks: an update. EXCLI journal 14, 346-378.

Lindeman, L.C., Andersen, I.S., Reiner, A.H., Li, N., Aanes, H., Ostrup, O., Winata, C., Mathavan, S., Muller, F., Alestrom, P., Collas, P., 2011. Prepatterning of developmental gene expression by modified histones before zygotic genome activation. Dev Cell 21, 993-1004.

Lippok, B., Song, S., Driever, W., 2014. Pou5f1 protein expression and posttranslational modification during early zebrafish development. Dev Dyn 243, 468-477.

Liu, Z., Li, W., Ma, X., Ding, N., Spallotta, F., Southon, E., Tessarollo, L., Gaetano, C., Mukouyama, Y.-s., Thiele, C.J., 2014. Essential Role of the Zinc Finger Transcription Factor Casz1 for Mammalian Cardiac Morphogenesis and Development. 289, 29801-29816.

Liu, Z., Lin, X., Cai, Z., Zhang, Z., Han, C., Jia, S., Meng, A., Wang, Q., 2011. Global identification of SMAD2 target genes reveals a role for multiple co-regulatory factors in zebrafish early gastrulas. J Biol Chem 286.

Liu, Z., Woo, S., Weiner, O.D., 2018. Nodal signaling has dual roles in fate specification and directed migration during germ layer segregation in zebrafish. Development 145.

Liu, Z.-P., 2015. Reverse Engineering of Genome-wide Gene Regulatory Networks from Gene Expression Data. Current genomics 16, 3-22.

Liu, Z.-P., Wu, C., Miao, H., Wu, H., 2015. RegNetwork: an integrated database of transcriptional and post-transcriptional regulatory networks in human and mouse. Database : the journal of biological databases and curation 2015, bav095.

Livak, K.J., Schmittgen, T.D., 2001. Analysis of relative gene expression data using real-time quantitative PCR and the 2(-Delta Delta C(T)) Method. Methods 25, 402-408.

Long, H.K., Prescott, S.L., Wysocka, J., 2016. Ever-Changing Landscapes: Transcriptional Enhancers in Development and Evolution. Cell 167, 1170-1187.

Long, S., Ahmad, N., Rebagliati, M., 2003. The zebrafish nodal-related gene southpaw is required for visceral and diencephalic left-right asymmetry. Development 130, 2303-2316.

Longabaugh, W.J., 2012. BioTapestry: a tool to visualize the dynamic properties of gene regulatory networks. Methods Mol Biol 786, 359-394.

Lou, X., Deshwar, A.R., Crump, J.G., Scott, I.C., 2011. Smarcd3b and Gata5 promote a cardiac progenitor fate in the zebrafish embryo. Development 138, 3113-3123.

Lough, J., Sugi, Y., 2000. Endoderm and heart development. Dev Dyn 217, 327-342.

Love, M.I., Huber, W., Anders, S., 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol 15.

Lovely, C.B., Swartz, M.E., McCarthy, N., Norrie, J.L., Eberhart, J.K., 2016. Bmp signaling mediates endoderm pouch morphogenesis by regulating Fgf signaling in zebrafish. Development 143, 2000-2011.

Lowe, E.K., Cuomo, C., Arnone, M.I., 2017. Omics approaches to study gene regulatory networks for development in echinoderms. Brief Funct Genomics 16, 299-308.

Lukoseviciute, M., Gavriouchkina, D., Williams, R.M., Hochgreb-Hagele, T., Senanayake, U., Chong-Morrison, V., Thongjuea, S., Repapi, E., Mead, A., Sauka-Spengler, T., 2018. From Pioneer to Repressor: Bimodal foxd3 Activity Dynamically Remodels Neural Crest Regulatory Landscape In Vivo. Developmental Cell 47, 608-628.e606.

Lunde, K., Belting, H.-G., Driever, W., 2004. Zebrafish pou5f1/pou2, homolog of mammalian Oct4, functions in the endoderm specification cascade. Current Biology 14, 48-55.

Ma, H., Lin, Y., Zhao, Z.-A., Lu, X., Yu, Y., Zhang, X., Wang, Q., Li, L., 2016. MicroRNA-127 Promotes Mesendoderm Differentiation of Mouse Embryonic Stem Cells by Targeting Left-Right Determination Factor 2. The Journal of biological chemistry 291, 12126-12135.

Macneil, L.T., Walhout, A.J.M., 2011. Gene regulatory networks and the role of robustness and stochasticity in the control of gene expression. Genome research 21, 645-657.

Maharana, S.K., Schlosser, G., 2018. A gene regulatory network underlying the formation of pre-placodal ectoderm in Xenopus laevis. BMC Biol 16, 79.

Mahony, S., Pugh, B.F., 2015. Protein-DNA binding in high-resolution. Crit Rev Biochem Mol Biol 50, 269-283.

Mangan, S., Alon, U., 2003. Structure and function of the feed-forward loop network motif. 100, 11980-11985.

Manning, A.J., Kimelman, D., 2015. Tbx16 and Msgn1 are required to establish directional cell migration of zebrafish mesodermal progenitors. Developmental Biology 406, 172-185.

Manoli, M., Driever, W., 2012. Fluorescence-activated cell sorting (FACS) of fluorescently tagged cells from zebrafish larvae for RNA isolation. Cold Spring Harb Protoc 2012.

Martik, M.L., Bronner, M.E., 2017. Regulatory Logic Underlying Diversification of the Neural Crest. Trends in Genetics 33, 715-727.

Maston, G.A., Evans, S.K., Green, M.R., 2006. Transcriptional regulatory elements in the human genome. Annual review of genomics and human genetics 7, 29-59.

Mathavan, S., Lee, S.G., Mak, A., Miller, L.D., Murthy, K.R., Govindarajan, K.R., Tong, Y., Wu, Y.L., Lam, S.H., Yang, H., 2005. Transcriptome analysis of zebrafish embryogenesis using microarrays. PLoS Genet 1.

Matsuda, K., Mikami, T., Oki, S., Iida, H., Andrabi, M., Boss, J.M., Yamaguchi, K., Shigenobu, S., Kondoh, H., 2017. ChIP-seq analysis of genomic binding regions of five major transcription factors highlights a central role for ZIC2 in the mouse epiblast stem cell gene regulatory network. Development 144, 1948-1958.

Matsui, T., Kanai-Azuma, M., Hara, K., Matoba, S., Hiramatsu, R., Kawakami, H., Kurohmaru, M., Koopman, P., Kanai, Y., 2006. Redundant roles of Sox17 and Sox18 in postnatal angiogenesis in mice. Journal of cell science 119, 3513-3526.

Mattick, J.S., Taft, R.J., Faulkner, G.J., 2010. A global view of genomic information--moving beyond the gene and the master regulator. Trends Genet 26, 21-28.

Matys, V., Kel-Margoulis, O.V., Fricke, E., Liebich, I., Land, S., Barre-Dirrie, A., Reuter, I., Chekmenev, D., Krull, M., Hornischer, K., Voss, N., Stegmaier, P., Lewicki-Potapov, B., Saxel, H., Kel, A.E., Wingender, E., 2006. TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. Nucleic Acids Res 34, D108-110.

Maves, L., Tyler, A., Moens, C.B., Tapscott, S.J., 2009. Pbx acts with Hand2 in early myocardial differentiation. Developmental Biology 333, 409-418.

Mazzoni, E.O., Mahony, S., Iacovino, M., Morrison, C.A., Mountoufaris, G., Closser, M., Whyte, W.A., Young, R.A., Kyba, M., Gifford, D.K., Wichterle, H., 2011. Embryonic stem cell-based system for mapping developmental transcriptional programs. Nature methods 8, 1056-1058.

McGettigan, P.A., 2013. Transcriptomics in the RNA-seq era. Current Opinion in Chemical Biology 17, 4-11.

McHaourab, Z.F., Perreault, A.A., Venters, B.J., 2018. ChIP-seq and ChIP-exo profiling of Pol II, H2A.Z, and H3K4me3 in human K562 cells. Sci Data 5, 180030.

Medina, I., Tárraga, J., Martínez, H., Barrachina, S., Castillo, M.I., Paschall, J., Salavert-Torres, J., Blanquer-Espert, I., Hernández-García, V., Quintana-Ortí, E.S., Dopazo, J., 2016. Highly sensitive and ultrafast read mapping for RNA-seq analysis. DNA research : an international journal for rapid publication of reports on genes and genomes 23, 93-100.

Meeker, N.D., Hutchinson, S.A., Ho, L., Trede, N.S., 2007. Method for isolation of PCR-ready genomic DNA from zebrafish tissues. BioTechniques 43, 610-614.

Melby, A.E., Warga, R.M., Kimmel, C.B., 1996. Specification of cell fates at the dorsal margin of the zebrafish gastrula. Development 122, 2225-2237.

Meno, C., Gritsman, K., Ohishi, S., Ohfuji, Y., Heckscher, E., Mochida, K., Shimono, A., Kondoh, H., Talbot, W.S., Robertson, E.J., Schier, A.F., Hamada, H., 1999. Mouse Lefty2 and zebrafish antivin are feedback inhibitors of nodal signaling during vertebrate gastrulation. Mol Cell 4, 287-298.

Meyer, C.A., Liu, X.S., 2014. Identifying and mitigating bias in next-generation sequencing methods for chromatin biology. Nat Rev Genet 15, 709-721.

Milo, R., Shen-Orr, S., Itzkovitz, S., Kashtan, N., Chklovskii, D., Alon, U., 2002. Network Motifs: Simple Building Blocks of Complex Networks. Science 298, 824.

Mitrophanov, A.Y., Groisman, E.A., 2008. Positive feedback in cellular control systems. BioEssays : news and reviews in molecular, cellular and developmental biology 30, 542-555.

Mizoguchi, T., Izawa, T., Kuroiwa, A., Kikuchi, Y., 2006. Fgf signaling negatively regulates Nodal-dependent endoderm induction in zebrafish. Developmental Biology 300, 612-622.

Mizoguchi, T., Verkade, H., Heath, J.K., Kuroiwa, A., Kikuchi, Y., 2008. Sdf1/Cxcr4 signaling controls the dorsal migration of endodermal cells during zebrafish gastrulation. Development 135, 2521.

Mohammadnia, A., Yaqubi, M., Pourasgari, F., Neely, E., Fallahi, H., Massumi, M., 2016. Signaling and Gene Regulatory Networks Governing Definitive Endoderm Derivation From Pluripotent Stem Cells. J Cell Physiol 231, 1994-2006.

Molkentin, J.D., 2000. The Zinc Finger-containing Transcription Factors GATA-4, -5, and -6: UBIQUITOUSLY EXPRESSED REGULATORS OF TISSUE-SPECIFIC GENE EXPRESSION. Journal of Biological Chemistry 275, 38949-38952.

Montague, T.G., Schier, A.F., 2017. Vg1-Nodal heterodimers are the endogenous inducers of mesendoderm. eLife 6, e28183.

Montero, J.-A., Carvalho, L., Wilsch-Bräuninger, M., Kilian, B., Mustafa, C., Heisenberg, C.-P., 2005. Shield formation at the onset of zebrafish gastrulation. Development 132, 1187.

Morley, R.H., Lachani, K., Keefe, D., Gilchrist, M.J., Flicek, P., Smith, J.C., Wardle, F.C., 2009. A gene regulatory network directed by zebrafish No tail accounts for its roles in mesoderm formation. Proceedings of the National Academy of Sciences 106, 3829-3834.

Mortazavi, A., Williams, B.A., McCue, K., Schaeffer, L., Wold, B., 2008. Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat Methods 5.

Mostafavi, S., Ortiz-Lopez, A., Bogue, M.A., Hattori, K., Pop, C., Koller, D., Mathis, D., Benoist, C., 2014. Variation and Genetic Control of Gene Expression in Primary Immunocytes across Inbred Mouse Strains. The Journal of Immunology 193, 4485.

Muruganujan, A., Ebert, D., Mi, H., Thomas, P.D., Huang, X., 2018. PANTHER version 14: more genomes, a new PANTHER GO-slim and improvements in enrichment analysis tools. Nucleic Acids Research 47, D419-D426.

Nagai, K., 2001. Molecular evolution of Sry and Sox gene. Gene 270, 161-169.

Nair, S., Schilling, T.F., 2008. Chemokine signaling controls endodermal migration during zebrafish gastrulation. Science (New York, N.Y.) 322, 89-92.

Nash, A.J., Tan, G., King, James W.D., Polychronopoulos, D., Lenhard, B., 2017. Conserved non-coding elements: developmental gene regulation meets genome organization. Nucleic Acids Research 45, 12611-12624.

Naval-Sánchez, M., Potier, D., Hulselmans, G., Christiaens, V., Aerts, S., 2015. Identification of Lineage-Specific Cis-Regulatory Modules Associated with Variation in Transcription Factor Binding and Chromatin Activity Using Ornstein-Uhlenbeck Models. Molecular biology and evolution 32, 2441-2455.

Nelson, A.C., Cutty, S.J., Gasiunas, S.N., Deplae, I., Stemple, D.L., Wardle, F.C., 2017. In Vivo Regulation of the Zebrafish Endoderm Progenitor Niche by T-Box Transcription Factors. Cell Rep 19, 2782-2795.

Nelson, A.C., Cutty, S.J., Niini, M., Stemple, D.L., Flicek, P., Houart, C., Bruce, A.E., Wardle, F.C., 2014. Global identification of Smad2 and Eomesodermin targets in zebrafish identifies a conserved transcriptional network in mesendoderm and a novel role for Eomesodermin in repression of ectodermal gene expression. BMC Biology 12, 1-20.

Nelson, A.C., Wardle, F.C., 2013. Conserved non-coding elements and cis regulation: actions speak louder than words. Development 140, 1385.

Nelson, C.J., Hurd, P.J., 2009. Advantages of next-generation sequencing versus the microarray in epigenetic research. Briefings in Functional Genomics 8, 174-183.

Nepal, C., Hadzhiev, Y., Previti, C., Haberle, V., Li, N., Takahashi, H., Suzuki, A.M.M., Sheng, Y., Abdelhamid, R.F., Anand, S., Gehrig, J., Akalin, A., Kockx, C.E.M., van der Sloot, A.A.J., van IJcken, W.F.J., Armant, O., Rastegar, S., Watson, C., Strähle, U., Stupka, E., Carninci, P., Lenhard, B., Müller, F., 2013. Dynamic regulation of the transcription initiation landscape at single nucleotide resolution during vertebrate embryogenesis. 23, 1938-1950.

Nevis, K., Obregon, P., Walsh, C., Guner-Ataman, B., Burns, C.G., Burns, C.E., 2013. Tbx1 is required for second heart field proliferation in zebrafish. Developmental dynamics : an official publication of the American Association of Anatomists 242, 550-559.

Niakan, K.K., Ji, H., Maehr, R., Vokes, S.A., Rodolfa, K.T., Sherwood, R.I., Yamaki, M., Dimos, J.T., Chen, A.E., Melton, D.A., McMahon, A.P., Eggan, K., 2010. Sox17 promotes differentiation in mouse embryonic stem cells by directly regulating extraembryonic gene expression and indirectly antagonizing self-renewal. Genes Dev 24, 312-326.

Nikolay Popgeorgiev, B.B., Julien Prudent and Germain Gillet, 2018. Control of Programmed Cell Death During Zebrafish Embryonic Development,. Recent Advances in Zebrafish Researches,.

Niwa, H., 2018. The principles that govern transcription factor network functions in stem cells. Development 145, dev157420.

Nowotschin, S., Costello, I., Piliszek, A., Kwon, G.S., Mao, C.-a., Klein, W.H., Robertson, E.J., Hadjantonakis, A.-K., 2013. The T-box transcription factor Eomesodermin is essential for AVE induction in the mouse embryo. Genes & development 27, 997-1002.

Nudelman, G., Frasca, A., Kent, B., Sadler, K.C., Sealfon, S.C., Walsh, M.J., Zaslavsky, E., 2018. High resolution annotation of zebrafish transcriptome using long-read sequencing. Genome Res 28, 1415-1425.

NuGEN, 2013. Detection of Genomic DNA in Human RNA Samples for RNA-Seq.

Ober, E.A., Field, H.A., Stainier, D.Y.R., 2003. From endoderm formation to liver and pancreas development in zebrafish. Mechanisms of development 120, 5-18.

Ober, E.A., Grapin-Botton, A., 2015. At new heights – endodermal lineages in development and disease. Development 142, 1912.

Odenthal, J., Nusslein-Volhard, C., 1998. fork head domain genes in zebrafish. Development genes and evolution 208, 245-258.

Okuda, Y., Yoda, H., Uchikawa, M., Furutani-Seiki, M., Takeda, H., Kondoh, H., Kamachi, Y., 2006. Comparative genomic and expression analysis of group B1 sox genes in zebrafish indicates their diversification during vertebrate evolution. Developmental Dynamics 235, 811-825.

Oliveri, P., Davidson, E.H., 2004. Gene regulatory network controlling embryonic specification in the sea urchin. Curr Opin Genet Dev 14, 351-360.

Osada, S.I., Wright, C.V., 1999. Xenopus nodal-related signaling is essential for mesendodermal patterning during early embryogenesis. Development 126, 3229-3240.

Osipovich, A.B., Long, Q., Manduchi, E., Gangula, R., Hipkens, S.B., Schneider, J., Okubo, T., Stoeckert, C.J., Takada, S., Magnuson, M.A., 2014. Insm1 promotes endocrine cell differentiation by modulating the expression of a network of genes that includes Neurog3 and Ripply3. Development 141, 2939.

Osterwalder, M., Barozzi, I., Tissieres, V., Fukuda-Yuzawa, Y., Mannion, B.J., Afzal, S.Y., Lee, E.A., Zhu, Y., Plajzer-Frick, I., Pickle, C.S., Kato, M., Garvin, T.H., Pham, Q.T., Harrington, A.N., Akiyama, J.A., Afzal, V., Lopez-Rios, J., Dickel, D.E., Visel, A., Pennacchio, L.A., 2018. Enhancer redundancy provides phenotypic robustness in mammalian development. Nature 554, 239-243.

Ott, E., Wendik, B., Srivastava, M., Pacho, F., Töchterle, S., Salvenmoser, W., Meyer, D., 2016. Pronephric tubule morphogenesis in zebrafish depends on Mnx mediated repression of irx1b within the intermediate mesoderm. Developmental Biology 411, 101-114.

Owens, N.D.L., Blitz, I.L., Lane, M.A., Patrushev, I., Overton, J.D., Gilchrist, M.J., Cho, K.W.Y., Khokha, M.K., 2016. Measuring Absolute RNA Copy Numbers at High Temporal Resolution Reveals Transcriptome Kinetics in Development. Cell Rep 14, 632-647.

Pai, A.A., Pritchard, J.K., Gilad, Y., 2015. The genetic and mechanistic basis for variation in gene regulation. PLoS Genet 11, e1004857.

Papasani, M.R., Robison, B.D., Hardy, R.W., Hill, R.A., 2006. Early developmental expression of two insulins in zebrafish (Danio rerio). Physiol Genomics 27, 79-85.

Parant, J.M., George, S.A., Pryor, R., Wittwer, C.T., Yost, H.J., 2009. A rapid and efficient method of genotyping zebrafish mutants. Developmental Dynamics 238, 3168-3174.

Parfitt, D.-E., Shen, M.M., 2014. From blastocyst to gastrula: gene regulatory networks of embryonic stem cells and early mouse embryogenesis. Philosophical transactions of the Royal Society of London. Series B, Biological sciences 369, 20130542.

Parsons, M.J., Pollard, S.M., Saúde, L., Feldman, B., Coutinho, P., Hirst, E.M.A., Stemple, D.L., 2002. Zebrafish mutants identify an essential role for laminins in notochord formation. Development 129, 3137.

Parvin, M.S., Okuyama, N., Inoue, F., Islam, M.E., Kawakami, A., Takeda, H., Yamasu, K., 2008. Autoregulatory loop and retinoic acid repression regulate pou2/pou5f1 gene expression in the zebrafish embryonic brain. Developmental Dynamics 237, 1373-1388.

Patro, R., Duggal, G., Love, M.I., Irizarry, R.A., Kingsford, C., 2017. Salmon provides fast and bias-aware quantification of transcript expression. Nature methods 14, 417-419.

Pauli, A., Valen, E., Lin, M.F., Garber, M., Vastenhouw, N.L., Levin, J.Z., Fan, L., Sandelin, A., Rinn, J.L., Regev, A., Schier, A.F., 2012. Systematic identification of long noncoding RNAs expressed during zebrafish embryogenesis. Genome Res 22, 577-591.

Pauls, S., Zecchin, E., Tiso, N., Bortolussi, M., Argenton, F., 2007. Function and regulation of zebrafish nkx2.2a during development of pancreatic islet and ducts. Developmental Biology 304, 875-890.

Pei, W., Noushmehr, H., Costa, J., Ouspenskaia, M.V., Elkahloun, A.G., Feldman, B., 2007. An early requirement for maternal FoxH1 during zebrafish gastrulation. Dev Biol 310.

Pendeville, H., Winandy, M., Manfroid, I., Nivelles, O., Motte, P., Pasque, V., Peers, B., Struman, I., Martial, J.A., Voz, M.L., 2008. Zebrafish Sox7 and Sox18 function together to control arterial–venous identity. Developmental Biology 317, 405-416.

Peng, Y., Yang, P.H., Ng, S.S., Lum, C.T., Kung, H.F., Lin, M.C., 2004. Protection of Xenopus laevis embryos against alcohol-induced delayed gut maturation and growth retardation by peroxiredoxin 5 and catalase. J Mol Biol 340, 819-827.

Pereira, L.A., Wong, M.S., Mei Lim, S., Stanley, E.G., Elefanty, A.G., 2012. The Mix family of homeobox genes—Key regulators of mesendoderm formation during vertebrate development. Developmental Biology 367, 163-177.

Perez-Camps, M., Tian, J., Chng, S.C., Sem, K.P., Sudhaharan, T., Teh, C., Wachsmuth, M., Korzh, V., Ahmed, S., Reversade, B., 2016. Quantitative imaging reveals real-time Pou5f3-Nanog complexes driving dorsoventral mesendoderm patterning in zebrafish. Elife 5.

Peter, I.S., 2017. Regulatory states in the developmental control of gene expression. Brief Funct Genomics 16, 281-287.

Peter, I.S., Davidson, E.H., 2010. The endoderm gene regulatory network in sea urchin embryos up to mid-blastula stage. Developmental Biology 340, 188-199.

Peter, Isabelle S., Davidson, Eric H., 2011a. Evolution of Gene Regulatory Networks Controlling Body Plan Development. Cell 144, 970-985.

Peter, I.S., Davidson, E.H., 2011b. A gene regulatory network controlling the embryonic specification of endoderm. Nature 474, 635-639.

Peterson, H., Reimand, J., Kolberg, L., Adler, P., Reisberg, S., Arak, T., Vilo, J., 2016. g:Profiler—a web server for functional interpretation of gene lists (2016 update). Nucleic Acids Research 44, W83-W89.

Peterson, S.M., Freeman, J.L., 2009. RNA isolation from embryonic zebrafish and cDNA synthesis for gene expression analysis. Journal of visualized experiments : JoVE, 1470.

Petratou, K., Subkhankulova, T., Lister, J.A., Rocco, A., Schwetlick, H., Kelsh, R.N., 2018. A systems biology approach uncovers the core gene regulatory network governing iridophore fate choice from the neural crest. PLOS Genetics 14, e1007402.

Peyrieras, N., Strahle, U., Rosa, F., 1998. Conversion of zebrafish blastomeres to an endodermal fate by TGF-beta-related signaling. Curr Biol 8, 783-786.

Pézeron, G., Mourrain, P., Courty, S., Ghislain, J., Becker, T.S., Rosa, F.M., David, N.B., 2008. Live Analysis of Endodermal Layer Formation Identifies Random Walk as a Novel Gastrulation Movement. Current Biology 18, 276-281.

Postlethwait, J.H., Woods, I.G., Ngo-Hazelett, P., Yan, Y.L., Kelly, P.D., Chu, F., Huang, H., Hill-Force, A., Talbot, W.S., 2000. Zebrafish comparative genomics and the origins of vertebrate chromosomes. Genome Res 10, 1890-1902.

Potier, D., Seyres, D., Guichard, C., Iche-Torres, M., Aerts, S., Herrmann, C., Perrin, L., 2014. Identification of cis-regulatory modules encoding temporal dynamics during development. BMC Genomics 15, 534.

Poulain, M., Furthauer, M., Thisse, B., Thisse, C., Lepage, T., 2006. Zebrafish endoderm formation is regulated by combinatorial Nodal, FGF and BMP signalling. Development 133, 2189-2200.

Poulain, M., Lepage, T., 2002. Mezzo, a paired-like homeobox protein is an immediate target of Nodal signalling and regulates endoderm specification in zebrafish. Development 129, 4901-4914.

Pownall, M.E., Gustafsson, M.K., Emerson, C.P., Jr., 2002. Myogenic regulatory factors and the specification of muscle progenitors in vertebrate embryos. Annu Rev Cell Dev Biol 18, 747-783.

Prior, H.M., Walter, M.A., 1996. SOX genes: architects of development. Molecular medicine (Cambridge, Mass.) 2, 405-412.

Prykhozhij, S.V., Marsico, A., Meijsing, S.H., 2013. Zebrafish Expression Ontology of Gene Sets (ZEOGS): a tool to analyze enrichment of zebrafish anatomical terms in large gene sets. Zebrafish 10, 303-315.

Pugh, B.F., Yamada, N., Farrell, N., Lai, W.K.M., Mahony, S., 2018. Characterizing protein–DNA binding event subtypes in ChIP-exo data. Bioinformatics 35, 903-913.

Qian, X., Ba, Y., Zhuang, Q., Zhong, G., 2014. RNA-Seq technology and its application in fish transcriptomics. OMICS 18, 98-110.

Qiao, L., Gao, H., Zhang, T., Jing, L., Xiao, C., Xiao, Y., Luo, N., Zhu, H., Meng, W., Xu, H., Mo, X., 2014. Snail modulates the assembly of fibronectin via α5 integrin for myocardial migration in zebrafish embryos. Scientific Reports 4, 4470.

Qu, X.B., Pan, J., Zhang, C., Huang, S.Y., 2008. Sox17 facilitates the differentiation of mouse embryonic stem cells into primitive and definitive endoderm in vitro. Dev Growth Differ 50, 585-593.

Quillien, A., Abdalla, M., Yu, J., Ou, J., Zhu, L.J., Lawson, N.D., 2017. Robust Identification of Developmentally Active Endothelial Enhancers in Zebrafish Using FANS-Assisted ATAC-Seq. Cell Rep 20, 709-720.

Quinlan, A.R., Hall, I.M., 2010. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26.

Rafiq, K., Shashikant, T., McManus, C.J., Ettensohn, C.A., 2014. Genome-wide analysis of the skeletogenic gene regulatory network of sea urchins. Development 141, 950.

Ramirez, F., Dundar, F., Diehl, S., Gruning, B.A., Manke, T., 2014. deepTools: a flexible platform for exploring deep-sequencing data. Nucleic Acids Res 42, W187-191.

Ramsköld, D., Luo, S., Wang, Y.C., Li, R., Deng, Q., Faridani, O.R., 2012. Full-length mRNA-Seq from single-cell levels of RNA and individual circulating tumor cells. Nat Biotechnol 30.

Rebagliati, M.R., Toyama, R., Haffter, P., Dawid, I.B., 1998. cyclops encodes a nodal-related factor involved in midline signaling. Proceedings of the National Academy of Sciences of the United States of America 95, 9932-9937.

Rehorn, K.P., Thelen, H., Michelson, A.M., Reuter, R., 1996. A molecular aspect of hematopoiesis and endoderm development common to vertebrates and Drosophila. Development 122, 4023-4031.

Reim, G., Mizoguchi, T., Stainier, D.Y., Kikuchi, Y., Brand, M., 2004. The POU Domain Protein Spg (Pou2/Oct4) Is Essential for Endoderm Formation in Cooperation with the HMG Domain Protein Casanova. Developmental Cell 6, 91-101.

Reiman, M., Laan, M., Rull, K., Sõber, S., 2017. Effects of RNA integrity on transcript quantification by total RNA sequencing of clinically collected human placental samples. The FASEB Journal 31, 3298-3308.

Reiter, F., Wienerroither, S., Stark, A., 2017. Combinatorial function of transcription factors and cofactors. Current Opinion in Genetics & Development 43, 73-81.

Reiter, J.F., Alexander, J., Rodaway, A., Yelon, D., Patient, R., Holder, N., Stainier, D.Y., 1999. Gata5 is required for the development of the heart and endoderm in zebrafish. Genes Dev 13.

Reiter, J.F., Kikuchi, Y., Stainier, D.Y., 2001. Multiple roles for Gata5 in zebrafish endoderm formation. Development 128.

Remenyi, A., Lins, K., Nissen, L.J., Reinbold, R., Scholer, H.R., Wilmanns, M., 2003. Crystal structure of a POU/HMG/DNA ternary complex suggests differential assembly of Oct4 and Sox2 on two enhancers. Genes Dev 17, 2048-2059.

Rex, M., Hilton, E., Old, R., 2002. Multiple interactions between maternally-activated signalling pathways control Xenopus nodal-related genes. Int J Dev Biol 46, 217-226.

Rhee, H.S., Pugh, B.F., 2012. ChIP-exo: A Method to Identify Genomic Location of DNA-binding proteins at Near Single Nucleotide Accuracy. Current protocols in molecular biology / edited by Frederick M. Ausubel ... [et al.] 0 21, 10.1002/0471142727.mb0471142124s0471142100.

Ridley, A.J., Paterson, H.F., Johnston, C.L., Diekmann, D., Hall, A., 1992. The small GTP-binding protein rac regulates growth factor-induced membrane ruffling. Cell 70, 401-410.

Rizzino, A., 2009. Sox2 and Oct-3/4: A versatile pair of master regulators that orchestrate the self-renewal and pluripotency of embryonic stem cells. Wiley interdisciplinary reviews. Systems biology and medicine 1, 228-236.

Robertson, E.J., 2014. Dose-dependent Nodal/Smad signals pattern the early mouse embryo. Semin Cell Dev Biol 32, 73-79.

Robinson, M.D., McCarthy, D.J., Smyth, G.K., 2009. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 26.

Rodaway, A., Takeda, H., Koshida, S., Broadbent, J., Price, B., Smith, J.C., Patient, R., Holder, N., 1999. Induction of the mesendoderm in the zebrafish germ ring by yolk cell-derived TGF-beta family signals and discrimination of mesoderm and endoderm by FGF. Development 126, 3067-3078.

Rogers, K.W., Lord, N.D., Gagnon, J.A., Pauli, A., Zimmerman, S., Aksel, D.C., Reyon, D., Tsai, S.Q., Joung, J.K., Schier, A.F., 2017. Nodal patterning without Lefty inhibitory feedback is functional but fragile. Elife 6.

Rojas, A., Schachterle, W., Xu, S.M., Martin, F., Black, B.L., 2010. Direct transcriptional regulation of Gata4 during early endoderm specification is controlled by FoxA2 binding to an intronic enhancer. Dev Biol 346, 346-355.

Rossi, A., Kontarakis, Z., Gerri, C., Nolte, H., Holper, S., Kruger, M., Stainier, D.Y., 2015. Genetic compensation induced by deleterious mutations but not gene knockdowns. Nature 524, 230-233.

Rossi, M.J., Lai, W.K.M., Pugh, B.F., 2018. Simplified ChIP-exo assays. Nature Communications 9, 2842.

Rougeot, J., Zakrzewska, A., Kanwal, Z., Jansen, H.J., Spaink, H.P., Meijer, A.H., 2014. RNA sequencing of FACS-sorted immune cell populations from zebrafish infection models to identify cell specific responses to intracellular pathogens. Methods Mol Biol 1197, 261-274.

Roy, S., Ernst, J., Kharchenko, P.V., Kheradpour, P., Negre, N., Eaton, M.L., Landolin, J.M., Bristow, C.A., Ma, L., Lin, M.F., Washietl, S., Arshinoff, B.I., Ay, F., Meyer, P.E., Robine, N., Washington, N.L., Di Stefano, L., Berezikov, E., Brown, C.D., Candeias, R., Carlson, J.W., Carr, A., Jungreis, I., Marbach, D., Sealfon, R., Tolstorukov, M.Y., Will, S., Alekseyenko, A.A., Artieri, C., Booth, B.W., Brooks, A.N., Dai, Q., Davis, C.A., Duff, M.O., Feng, X., Gorchakov, A.A., Gu, T., Henikoff, J.G., Kapranov, P., Li, R., MacAlpine, H.K., Malone, J., Minoda, A., Nordman, J., Okamura, K., Perry, M., Powell, S.K., Riddle, N.C., Sakai, A., Samsonova, A., Sandler, J.E., Schwartz, Y.B., Sher, N., Spokony, R., Sturgill, D., van Baren, M., Wan, K.H., Yang, L., Yu, C., Feingold, E., Good, P., Guyer, M., Lowdon, R., Ahmad, K., Andrews, J., Berger, B., Brenner, S.E., Brent, M.R., Cherbas, L., Elgin, S.C., Gingeras, T.R., Grossman, R., Hoskins, R.A., Kaufman, T.C., Kent, W., Kuroda, M.I., Orr-Weaver, T., Perrimon, N., Pirrotta, V., Posakony, J.W., Ren, B., Russell, S., Cherbas, P., Graveley, B.R., Lewis, S., Micklem, G., Oliver, B., Park, P.J., Celniker, S.E., Henikoff, S., Karpen, G.H., Lai, E.C., MacAlpine, D.M., Stein, L.D., White, K.P., Kellis, M., 2010. Identification of functional elements and regulatory circuits by Drosophila modENCODE. Science 330, 1787-1797.

Roy, S., Raj, M., Ghosh, P., Das, S.K., 2017. Role of motifs in topological robustness of gene regulatory networks, 2017 IEEE International Conference on Communications (ICC), pp. 1-6.

Ruprecht, V., Wieser, S., Callan-Jones, A., Smutny, M., Morita, H., Sako, K., Barone, V., Ritsch-Marte, M., Sixt, M., Voituriez, R., Heisenberg, C.-P., 2015. Cortical Contractility Triggers a Stochastic Switch to Fast Amoeboid Cell Motility. Cell 160, 673-685.

Ruzicka, L., Bradford, Y.M., Frazer, K., Howe, D.G., Paddock, H., Ramachandran, S., Singer, A., Toro, S., Van Slyke, C.E., Eagle, A.E., Fashena, D., Kalita, P., Knight, J., Mani, P., Martin, R., Moxon, S.A.T., Pich, C., Schaper, K., Shao, X., Westerfield, M., 2015. ZFIN, The zebrafish model organism database: Updates and new directions. genesis 53, 498-509.

Sahraeian, S.M.E., Mohiyuddin, M., Sebra, R., Tilgner, H., Afshar, P.T., Au, K.F., Bani Asadi, N., Gerstein, M.B., Wong, W.H., Snyder, M.P., Schadt, E., Lam, H.Y.K., 2017. Gaining comprehensive biological insight into the transcriptome by performing a broad-spectrum RNA-seq analysis. Nature Communications 8, 59.

Sakaguchi, T., Kikuchi, Y., Kuroiwa, A., Takeda, H., Stainier, D.Y.R., 2006. The yolk syncytial layer regulates myocardial migration by influencing extracellular matrix assembly in zebrafish. Development 133, 4063.

Sakaguchi, T., Kuroiwa, A., Takeda, H., 2001. A novel sox gene, 226D7, acts downstream of Nodal signaling to specify endoderm precursors in zebrafish. Mechanisms of development 107, 25-38.

Sako, K., Pradhan, S.J., Barone, V., Ingles-Prieto, A., Muller, P., Ruprecht, V., Capek, D., Galande, S., Janovjak, H., Heisenberg, C.P., 2016. Optogenetic Control of Nodal Signaling Reveals a Temporal Pattern of Nodal Signaling Regulating Cell Fate Specification during Gastrulation. Cell Rep 16, 866-877.

Salazar-Ciudad, I., Jernvall, J., Newman, S.A., 2003. Mechanisms of pattern formation in development and evolution. Development 130, 2027.

Sampath, K., Rubinstein, A.L., Cheng, A.M., Liang, J.O., Fekany, K., Solnica-Krezel, L., Korzh, V., Halpern, M.E., Wright, C.V., 1998. Induction of the zebrafish ventral brain and floorplate requires cyclops/nodal signalling. Nature 395.

Sanchita, Sharma, A., 2015. In silico identification of regulatory motifs in co-expressed genes under osmotic stress representing their co-regulation. Plant Gene 1, 29-34.

Sarropoulou, E., Galindo-Villegas, J., Garcia-Alcazar, A., Kasapidis, P., Mulero, V., 2012. Characterization of European sea bass transcripts by RNA SEQ after oral vaccine against V. anguillarum. Mar Biotechnol (NY) 14, 634-642.

Satou, Y., Imai, K.S., 2015. Gene regulatory systems that control gene expression in the Ciona embryo. Proceedings of the Japan Academy. Series B, Physical and biological sciences 91, 33-51.

Savic, D., Partridge, E.C., Newberry, K.M., Smith, S.B., Meadows, S.K., Roberts, B.S., Mackiewicz, M., Mendenhall, E.M., Myers, R.M., 2015. CETCh-seq: CRISPR epitope tagging ChIP-seq of DNA-binding proteins. Genome research 25, 1581-1589.

Schepis, A., Nelson, W.J., 2012. Adherens junction function and regulation during zebrafish gastrulation. Cell adhesion & migration 6, 173-178.

Schier, A.F., 2003. Nodal Signaling in Vertebrate Development. Annual Review of Cell and Developmental Biology 19, 589-621.

Schier, A.F., Neuhauss, S.C., Helde, K.A., Talbot, W.S., Driever, W., 1997. The one-eyed pinhead gene functions in mesoderm and endoderm formation in zebrafish and interacts with no tail. Development 124, 327-342.

Schier, A.F., Shen, M.M., 2000. Nodal signalling in vertebrate development. Nature 403, 385-389.

Schier, A.F., Talbot, W.S., 2005. Molecular Genetics of Axis Formation in Zebrafish. Annual Review of Genetics 39, 561-613.

Schilders, K., Ochieng, J.K., van de Ven, C.P., Gontan, C., Tibboel, D., Rottier, R.J., 2014. Role of SOX2 in foregut development in relation to congenital abnormalities. . World J Med Genet 4, 94-104.

Schindelin, J., Arganda-Carreras, I., Frise, E., Kaynig, V., Longair, M., Pietzsch, T., Preibisch, S., Rueden, C., Saalfeld, S., Schmid, B., Tinevez, J.Y., White, D.J., Hartenstein, V., Eliceiri, K., Tomancak, P., Cardona, A., 2012. Fiji: an open-source platform for biological-image analysis. Nat Methods 9, 676-682.

Schmidt, D., Wilson, M.D., Spyrou, C., Brown, G.D., Hadfield, J., Odom, D.T., 2009. ChIP-seq: using high-throughput sequencing to discover protein-DNA interactions. Methods 48.

Schottenfeld, J., Sullivan-Brown, J., Burdine, R.D., 2007. Zebrafish curly up encodes a Pkd2 ortholog that restricts left-side-specific expression of southpaw. Development 134, 1605-1615.

Schulze, S.K., Kanwar, R., Gölzenleuchter, M., Therneau, T.M., Beutler, A.S.J.B.G., 2012. SERE: Single-parameter quality control and sample comparison for RNA-Seq. 13, 524.

Schurch, N.J., Schofield, P., Gierlinski, M., Cole, C., Sherstnev, A., Singh, V., Wrobel, N., Gharbi, K., Simpson, G.G., Owen-Hughes, T., Blaxter, M., Barton, G.J., 2016. How many biological replicates are needed in an RNA-seq experiment and which differential expression tool should you use? Rna 22, 839-851.

Schwartz, S., Oren, R., Ast, G., 2011. Detection and removal of biases in the analysis of next-generation sequencing reads. PloS one 6, e16685-e16685.

Serandour, A.A., Brown, G.D., Cohen, J.D., Carroll, J.S., 2013. Development of an Illumina-based ChIP-exonuclease method provides insight into FoxA1-DNA binding properties. Genome Biology 14, 1-9.

Seyednasrollah, F., Laiho, A., Elo, L.L., 2015. Comparison of software packages for detecting differential expression in RNA-seq studies. Brief Bioinform 16.

Sharifi-Zarchi, A., Totonchi, M., Khaloughi, K., Karamzadeh, R., Arauzo-Bravo, M.J., Baharvand, H., Tusserkani, R., Pezeshk, H., Chitsaz, H., Sadeghi, M., 2015. Increased robustness of early embryogenesis through collective decision-making by key transcription factors. BMC Syst Biol 9, 23.

She, Z.-Y., Yang, W.-X., 2015. SOX family transcription factors involved in diverse cellular events during development. European Journal of Cell Biology 94, 547-563.

Shen, M.M., 2007. Nodal signaling: developmental roles and regulation. Development 134, 1023-1034.

Shen, N., Zhao, J., Schipper, J.L., Zhang, Y., Bepler, T., Leehr, D., Bradley, J., Horton, J., Lapp, H., Gordan, R., 2018. Divergence in DNA Specificity among Paralogous Transcription Factors Contributes to Their Differential In Vivo Binding. Cell Systems 6, 470-483.e478.

Sheng, Q., Vickers, K., Zhao, S., Wang, J., Samuels, D.C., Koues, O., Shyr, Y., Guo, Y., 2017. Multi-perspective quality control of Illumina RNA sequencing data analysis. Brief Funct Genomics 16, 194-204.

Shi, W., Levine, M., Davidson, B., 2005. Unraveling genomic regulatory networks in the simple chordate, Ciona intestinalis. Genome Res 15, 1668-1674.

Shimomura, O., Johnson, F.H., Saiga, Y., 1962. Extraction, Purification and Properties of Aequorin, a Bioluminescent Protein from the Luminous Hydromedusan, Aequorea. Journal of Cellular and Comparative Physiology 59, 223-239.

Simões-Costa, M., Bronner, M.E., 2015. Establishing neural crest identity: a gene regulatory recipe. Development 142, 242.

Simoes-Costa, M., Tan-Cabugao, J., Antoshechkin, I., Sauka-Spengler, T., Bronner, M.E., 2014. Transcriptome analysis reveals novel players in the cranial neural crest gene regulatory network. Genome Res 24, 281-290.

Singh, A.J., Chang, C.-N., Ma, H.-Y., Ramsey, S.A., Filtz, T.M., Kioussi, C., 2018. FACS-Seq analysis of Pax3-derived cells identifies non-myogenic lineages in the embryonic forelimb. Scientific Reports 8, 7670.

Singh, A.R., Sivadas, A., Sabharwal, A., Vellarikal, S.K., Jayarajan, R., Verma, A., Kapoor, S., Joshi, A., Scaria, V., Sivasubbu, S., 2016. Chamber Specific Gene Expression Landscape of the Zebrafish Heart. PloS one 11, e0147823-e0147823.

Singh, H., Khan, A.A., Dinner, A.R., 2014. Gene regulatory networks in the immune system. Trends in immunology 35, 211-218.

Singh, V.K., Kalsan, M., Kumar, N., Saini, A., Chandra, R., 2015. Induced pluripotent stem cells: applications in regenerative medicine, disease modeling, and drug discovery. Frontiers in cell and developmental biology 3, 2-2.

Sinner, D., Kirilenko, P., Rankin, S., Wei, E., Howard, L., Kofron, M., Heasman, J., Woodland, H.R., Zorn, A.M., 2006. Global analysis of the transcriptional network controlling Xenopus endoderm formation. Development 133, 1955.

Sinner, D., Rankin, S., Lee, M., Zorn, A.M., 2004a. Sox17 and beta-catenin cooperate to regulate the transcription of endodermal genes. Development 131, 3069-3080.

Sinner, D., Rankin, S., Lee, M., Zorn, A.M., 2004b. Sox17 and β-catenin cooperate to regulate the transcription of endodermal genes. Development 131, 3069.

Slagle, C.E., Aoki, T., Burdine, R.D., 2011. Nodal-dependent mesendoderm specification requires the combinatorial activities of FoxH1 and Eomesodermin. PLoS Genet 7.

Smith, S., Bernatchez, L., Beheregaray, L.B., 2013. RNA-seq analysis reveals extensive transcriptional plasticity to temperature stress in a freshwater fish species. BMC genomics 14, 375-375.

Smyth, G.K., 2005. limma: Linear Models for Microarray Data, in: Gentleman, R., Carey, V.J., Huber, W., Irizarry, R.A., Dudoit, S. (Eds.), Bioinformatics and Computational Biology Solutions Using R and Bioconductor. Springer New York, New York, NY, pp. 397-420.

Smyth, N., Vatansever, H.S., Murray, P., Meyer, M., Frie, C., Paulsson, M., Edgar, D., 1999. Absence of Basement Membranes after Targeting the LAMC1 Gene Results in Embryonic Lethality Due to Failure of Endoderm Differentiation The Journal of Cell Biology 144, 151.

Sojka, S., Amin, N.M., Gibbs, D., Christine, K.S., Charpentier, M.S., Conlon, F.L., 2014. Congenital heart disease protein 5 associates with CASZ1 to maintain myocardial tissue integrity. Development (Cambridge, England) 141, 3040-3049.

Solnica-Krezel, L., 2002. Pattern Formation in Zebrafish. . Springer-Verlag, 438 p.

Solnica-Krezel, L., Stemple, D.L., Mountcastle-Shah, E., Rangini, Z., Neuhauss, S.C., Malicki, J., Schier, A.F., Stainier, D.Y., Zwartkruis, F., Abdelilah, S., Driever, W., 1996. Mutations

affecting cell fates and cellular rearrangements during gastrulation in zebrafish. Development 123.

Solomon, K.S., Kudoh, T., Dawid, I.B., Fritz, A., 2003. Zebrafish foxi1 mediates otic placode formation and jaw development. Development 130, 929.

Spitz, F., Furlong, E.E.M., 2012. Transcription factors: from enhancer binding to developmental control. Nat Rev Genet 13, 613-626.

Stainier, D.Y., Fouquet, B., Chen, J.N., Warren, K.S., Weinstein, B.M., Meiler, S.E., Mohideen, M.A., Neuhauss, S.C., Solnica-Krezel, L., Schier, A.F., Zwartkruis, F., Stemple, D.L., Malicki, J., Driever, W., Fishman, M.C., 1996. Mutations affecting the formation and function of the cardiovascular system in the zebrafish embryo. Development 123, 285-292.

Stainier, D.Y.R., 2002. A glimpse into the molecular entrails of endoderm formation. Genes & Development 16, 893-907.

Stainier, D.Y.R., Raz, E., Lawson, N.D., Ekker, S.C., Burdine, R.D., Eisen, J.S., Ingham, P.W., Schulte-Merker, S., Yelon, D., Weinstein, B.M., Mullins, M.C., Wilson, S.W., Ramakrishnan, L., Amacher, S.L., Neuhauss, S.C.F., Meng, A., Mochizuki, N., Panula, P., Moens, C.B., 2017. Guidelines for morpholino use in zebrafish. PLOS Genetics 13, e1007000.

Starick, S.R., Ibn-Salem, J., Jurk, M., Hernandez, C., Love, M.I., Chung, H.R., Vingron, M., Thomas-Chollier, M., Meijsing, S.H., 2015. ChIP-exo signal associated with DNA-binding motifs provides insight into the genomic binding of the glucocorticoid receptor and cooperating transcription factors. Genome Res 25, 825-835.

Stickney, H.L., Imai, Y., Draper, B., Moens, C., Talbot, W.S., 2007. Zebrafish bmp4 functions during late gastrulation to specify ventroposterior cell fates. Developmental biology 310, 71-84.

Stolt, C.C., Schlierf, A., Lommes, P., Hillgärtner, S., Werner, T., Kosian, T., Sock, E., Kessaris, N., Richardson, W.D., Lefebvre, V., Wegner, M., 2006. SoxD Proteins Influence Multiple Stages of Oligodendrocyte Development and Modulate SoxE Protein Function. Developmental Cell 11, 697-709.

Strahle, U., Blader, P., Henrique, D., Ingham, P.W., 1993. Axial, a zebrafish gene expressed along the developing body axis, shows altered expression in cyclops mutant embryos. Genes Dev 7, 1436-1446.

Straub, B.K., Rickelt, S., Zimbelmann, R., Grund, C., Kuhn, C., Iken, M., Ott, M., Schirmacher, P., Franke, W.W., 2011. E–N-cadherin heterodimers define novel adherens junctions connecting endoderm-derived cells. The Journal of Cell Biology 195, 873.

Struckmann, S., Esch, D., Schöler, H., Fuellen, G., 2011. Visualization and exploration of conserved regulatory modules using ReXSpecies 2. BMC evolutionary biology 11, 267-267.

Stuart, G.W., McMurray, J.V., Westerfield, M., 1988. Replication, integration and stable germ-line transmission of foreign sequences injected into early zebrafish embryos. Development 103, 403-412.

Stuart, G.W., Vielkind, J.R., McMurray, J.V., Westerfield, M., 1990. Stable lines of transgenic zebrafish exhibit reproducible patterns of transgene expression. Development 109, 577-584.

Sultan, M., Amstislavskiy, V., Risch, T., Schuette, M., Dökel, S., Ralser, M., Balzereit, D., Lehrach, H., Yaspo, M.-L., 2014. Influence of RNA extraction methods and library selection schemes on RNA-seq data. BMC genomics 15, 675-675.

Svensson, V., Teichmann, S.A., Stegle, O., 2018. SpatialDE: identification of spatially variable genes. Nature Methods 15, 343.

Swift, M.R., Pham, V.N., Castranova, D., Bell, K., Poole, R.J., Weinstein, B.M., 2014. SoxF factors and Notch regulate nr2f2 gene expression during venous differentiation in zebrafish. Dev Biol 390, 116-125.

Takada, S., Okada, K., Wada, H., 2018. Endoderm morphogenesis reveals integration of distinct processes in the development and evolution of pharyngeal arches. biorxiv.

Takizawa, F., Araki, K., Ito, K., Moritomo, T., Nakanishi, T., 2007. Expression analysis of two Eomesodermin homologues in zebrafish lymphoid tissues and cells. Mol Immunol 44.

Talbot, J.C., Johnson, S.L., Kimmel, C.B., 2010. hand2 and Dlx genes specify dorsal, intermediate and ventral domains within zebrafish pharyngeal arches. Development 137, 2507.

Talbot, J.C., Walker, M.B., Carney, T.J., Huycke, T.R., Yan, Y.-L., BreMiller, R.A., Gai, L., Delaurier, A., Postlethwait, J.H., Hammerschmidt, M., Kimmel, C.B., 2012. fras1 shapes endodermal pouch 1 and stabilizes zebrafish pharyngeal skeletal development. Development (Cambridge, England) 139, 2804-2813.

Tam, P.P.L., Loebel, D.A.F., 2007. Gene function in mouse embryogenesis: get set for gastrulation. Nature Reviews Genetics 8, 368.

Tan, H., Onichtchouk, D., Winata, C., 2016. DANIO-CODE: Toward an Encyclopedia of DNA Elements in Zebrafish. Zebrafish 13, 54-60.

Tang, R., Dodd, A., Lai, D., McNabb, W.C., Love, D.R., 2007. Validation of Zebrafish (Danio rerio) Reference Genes for Quantitative Real-time RT-PCR Normalization. Acta Biochimica et Biophysica Sinica 39, 384-390.

Tang, X., Liu, H., Srivastava, A., Pécot, T., Chen, Z., Wang, Q., Huang, K., Sáenz-Robles, M.T., Cantalupo, P., Pipas, J., Leone, G., 2016. Transcriptome regulation and chromatin occupancy by E2F3 and MYC in mice. Scientific Data 3, 160008.

Tarazona, S., Furió-Tarí, P., Turrà, D., Pietro, A.D., Nueda, M.J., Ferrer, A., 2015. Data quality aware analysis of differential expression in RNA-seq with NOISeq R/Bioc package. Nucleic Acids Res 43.

Tarifeño-Saldivia, E., Lavergne, A., Bernard, A., Padamata, K., Bergemann, D., Voz, M.L., Manfroid, I., Peers, B., 2017. Transcriptome analysis of pancreatic cells across distant species highlights novel important regulator genes. BMC biology 15, 21-21.

Teo, A.K., Arnold, S.J., Trotter, M.W., Brown, S., Ang, L.T., Chng, Z., Robertson, E.J., Dunn, N.R., Vallier, L., 2011. Pluripotency factors regulate definitive endoderm specification through eomesodermin. Genes Dev 25.

Thisse, B., Wright, C.V.E., Thisse, C., 2000. Activin- and Nodal-related factors control antero–posterior patterning of the zebrafish embryo. Nature 403, 425-428.

Thisse, C., Thisse, B., 2008. High-resolution in situ hybridization to whole-mount zebrafish embryos. Nat Protoc 3, 59-69.

Tian, T., Zhao, L., Zhang, M., Zhao, X., Meng, A., 2009. Both foxj1a and foxj1b are implicated in left-right asymmetric development in zebrafish embryos. Biochemical and biophysical research communications 380, 537-542.

Trinh, L.A., Chong-Morrison, V., Gavriouchkina, D., Hochgreb-Hagele, T., Senanayake, U., Fraser, S.E., Sauka-Spengler, T., 2017. Biotagging of Specific Cell Populations in Zebrafish Reveals Gene Regulatory Logic Encoded in the Nuclear Transcriptome. Cell Rep 19, 425-440.

Tseng, W.F., Jang, T.H., Huang, C.B., Yuh, C.H., 2011. An evolutionarily conserved kernel of gata5, gata6, otx2 and prdm1a operates in the formation of endoderm in zebrafish. Dev Biol 357, 541-557.

Vallier, L., Mendjan, S., Brown, S., Chng, Z., Teo, A., Smithers, L.E., Trotter, M.W.B., Cho, C.H.H., Martinez, A., Rugg-Gunn, P., Brons, G., Pedersen, R.A., 2009. Activin/Nodal signalling maintains pluripotency by controlling Nanog expression. Development (Cambridge, England) 136, 1339-1349.

van Boxtel, A.L., Chesebro, J.E., Heliot, C., Ramel, M.C., Stone, R.K., Hill, C.S., 2015. A Temporal Window for Signal Activation Dictates the Dimensions of a Nodal Signaling Domain. Dev Cell 35, 175-185.

van Boxtel, A.L., Economou, A.D., Heliot, C., Hill, C.S., 2018. Long-Range Signaling Activation and Local Inhibition Separate the Mesoderm and Endoderm Lineages. Dev Cell 44, 179-191.e175.

van der Graaf, A., Franke, L., Võsa, U., van Dam, S., de Magalhães, J.P., 2017. Gene co-expression analysis for functional classification and gene–disease predictions. Briefings in Bioinformatics 19, 575-592.

Van Peer, G., Mestdagh, P., Vandesompele, J., 2012. Accurate RT-qPCR gene expression analysis on cell culture lysates. Scientific reports 2, 222-222.

Varlet, I., Collignon, J., Robertson, E.J., 1997. Nodal expression in the primitive endoderm is required for specification of the anterior axis during mouse gastrulation. Development 124.

Veerla, S., Höglund, M., 2006. Analysis of promoter regions of co-expressed genes identified by microarray analysis. BMC bioinformatics 7, 384-384.

Veil, M., Schaechtle, M.A., Gao, M., Kirner, V., Buryanova, L., Grethen, R., Onichtchouk, D., 2018. Maternal Nanog is required for zebrafish embryo architecture and for cell viability during gastrulation. Development 145, dev155366.

Vesterlund, L., Jiao, H., Unneberg, P., Hovatta, O., Kere, J., 2011. The zebrafish transcriptome during early development. BMC Dev Biol 11, 30.

Vignali, R., Poggi, L., Madeddu, F., Barsacchi, G., 2000. HNF1(beta) is required for mesoderm induction in the Xenopus embryo. Development 127, 1455-1465.

Vogan, K., 2015. Zebrafish mutants versus morphants. Nature Genetics 47, 105.

Voldoire, E., Brunet, F., Naville, M., Volff, J.-N., Galiana, D., 2017. Expansion by whole genome duplication and evolution of the sox gene family in teleost fish. PLOS ONE 12, e0180936.

Vopalensky, P., Pralow, S., Vastenhouw, N.L., 2018. Reduced expression of the Nodal co-receptor Oep causes loss of mesendodermal competence in zebrafish. Development 145.

Voronina, A., Pshennikova, E., 2016. The Vox mRNA and protein expression in zebrafish Pou5f3 MZspg mutant embryos. Stem Cell Investig 3, 79.

Waddington, C.H., 1940. Organisers & genes. Cambridge University Press Cambridge

Wagner, A., 2008. Gene duplications, robustness and evolutionary innovations. Bioessays 30, 367-373.

Wagner, D.E., Weinreb, C., Collins, Z.M., Briggs, J.A., Megason, S.G., Klein, A.M., 2018. Single-cell mapping of gene expression landscapes and lineage in the zebrafish embryo. Science 360, 981.

Wallace, K.N., Yusuff, S., Sonntag, J.M., Chin, A.J., Pack, M., 2001. Zebrafish hhex regulates liver development and digestive organ chirality. Genesis 30, 141-143.

Wang, L., Chen, J., Wang, C., Uusküla-Reimand, L., Chen, K., Medina-Rivera, A., Young, E.J., Zimmermann, M.T., Yan, H., Sun, Z., Zhang, Y., Wu, S.T., Huang, H., Wilson, M.D., Kocher, J.-P.A., Li, W., 2014. MACE: model based analysis of ChIP-exo. Nucleic Acids Research.

Wang, L., Felts, S.J., Van Keulen, V.P., Pease, L.R., Zhang, Y., 2018. Exploring the effect of library preparation on RNA sequencing experiments. Genomics.

Wang, L., Nie, J., Sicotte, H., Li, Y., Eckel-Passow, J.E., Dasari, S., Vedell, P.T., Barman, P., Wang, L., Weinshiboum, R., Jen, J., Huang, H., Kohli, M., Kocher, J.-P.A., 2016. Measure transcript integrity using RNA-seq data. BMC bioinformatics 17, 58-58.

Wang, L., Wang, S., Li, W., 2012. RSeQC: quality control of RNA-seq experiments. Bioinformatics 28.

Wardle, F.C., Tan, H., 2015. A ChIP on the shoulder? Chromatin immunoprecipitation and validation strategies for ChIP antibodies. F1000Research 4.

Warga, R.M., Kane, D.A., 2007. A role for N-cadherin in mesodermal morphogenesis during gastrulation. Developmental Biology 310, 211-225.

Warga, R.M., Kimmel, C.B., 1990. Cell movements during epiboly and gastrulation in zebrafish. Development 108, 569-580.

Warga, R.M., Nusslein-Volhard, C., 1999. Origin and development of the zebrafish endoderm. Development 126, 827-838.

Webb, A.E., Sanderford, J., Frank, D., Talbot, W.S., Driever, W., Kimelman, D., 2007. Laminin alpha5 is essential for the formation of the zebrafish fins. Dev Biol 311, 369-382.

Weber, H., Symes, C.E., Walmsley, M.E., Rodaway, A.R., Patient, R.K., 2000. A role for GATA5 in Xenopus endoderm specification. Development 127, 4345-4360.

Wegner, M., 2010. All purpose Sox: The many roles of Sox proteins in gene expression. Int J Biochem Cell Biol 42, 381-390.

Wen, B., Yuan, H., Liu, X., Wang, H., Chen, S., Chen, Z., de The, H., Zhou, J., Zhu, J., 2017. GATA5 SUMOylation is indispensable for zebrafish cardiac development. Biochimica et biophysica acta. General subjects 1861, 1691-1701.

Weng, W., Stemple, D.L., 2003. Nodal signaling and vertebrate germ layer formation. Birth Defects Res C Embryo Today 69.

White, R.J., Collins, J.E., Sealy, I.M., Wali, N., Dooley, C.M., Digby, Z., Stemple, D.L., Murphy, D.N., Billis, K., Hourlier, T., Fullgrabe, A., Davis, M.P., Enright, A.J., Busch-Nentwich, E.M., 2017. A high-resolution mRNA expression time course of embryonic development in zebrafish. Elife 6.

Wilfinger, A., Arkhipova, V., Meyer, D., 2013. Cell type and tissue specific function of islet genes in zebrafish pancreas development. Dev Biol 378, 25-37.

Williams, R.M., Candido-Ferreira, I., Repapi, E., Gavriouchkina, D., Senanayake, U., Telenius, J., Taylor, S., Hughes, J., Sauka-Spengler, T., 2018. Reconstruction of the global neural crest gene regulatory network in vivo. bioRxiv, 508473.

Wilson, M.J., Dearden, P.K., 2008. Evolution of the insect Sox genes. BMC evolutionary biology 8, 120.

Wimmer, I., Tröscher, A.R., Brunner, F., Rubino, S.J., Bien, C.G., Weiner, H.L., Lassmann, H., Bauer, J., 2018. Systematic evaluation of RNA quality, microarray data reliability and pathway analysis in fresh, fresh frozen and formalin-fixed paraffin-embedded tissue samples. Scientific Reports 8, 6351.

Winata, C.L., Kondrychyn, I., Kumar, V., Srinivasan, K.G., Orlov, Y., Ravishankar, A., Prabhakar, S., Stanton, L.W., Korzh, V., Mathavan, S., 2013. Genome wide analysis reveals Zic3 interaction with distal regulatory elements of stage specific developmental genes in zebrafish. PLoS Genet 9, e1003852.

Winkler, C., Schäfer, M., Duschl, J., Schartl, M., Volff, J.-N., 2003. Functional Divergence of Two Zebrafish Midkine Growth Factors Following Fish-Specific Gene Duplication. 13, 1067-1081.

Wissmuller, S., Kosian, T., Wolf, M., Finzsch, M., Wegner, M., 2006. The high-mobility-group domain of Sox proteins interacts with DNA-binding domains of many transcription factors. Nucleic Acids Res 34, 1735-1744.

Wittwer, C.T., Palais, R., Dwight, Z., 2011. uMELT: prediction of high-resolution melting curves and dynamic melting profiles of PCR products in a rich web application. Bioinformatics 27, 1019-1020.

Witzel, H.R., Cheedipudi, S., Gao, R., Stainier, D.Y.R., Dobreva, G.D., 2017. Isl2b regulates anterior second heart field development in zebrafish. Scientific Reports 7, 41043.

Witzel, H.R., Jungblut, B., Choe, C.P., Crump, J.G., Braun, T., Dobreva, G., 2012. The LIM protein Ajuba restricts the second heart field progenitor pool by regulating Isl1 activity. Dev Cell 23, 58-70.

Wong, W., Farr, R., Joglekar, M., Januszewski, A., Hardikar, A., 2015. Probe-based Real-time PCR Approaches for Quantitative Measurement of microRNAs. J Vis Exp.

Woo, S., Housley, M.P., Weiner, O.D., Stainier, D.Y.R., 2012. Nodal signaling regulates endodermal cell motility and actin dynamics via Rac1 and Prex1. The Journal of Cell Biology 198, 941.

Worsley Hunt, R., Wasserman, W.W., 2014. Non-targeted transcription factors motifs are a systemic component of ChIP-seq datasets. Genome Biol 15, 412.

Wu, D., Zhao, L., Skrbo-Larssen, N., Yang, S., Tian, T., Zheng, X., Zhao, X., Han, Y., Kuang, Z., Meng, A., Zhang, C., Lu, Q., 2008. Heart-specific isoform of tropomyosin4 is essential for heartbeat in zebrafish embryos. Cardiovascular Research 80, 200-208.

Xing, L., Quist, T.S., Stevenson, T.J., Dahlem, T.J., Bonkowsky, J.L., 2014. Rapid and efficient zebrafish genotyping using PCR with high-resolution melt analysis. Journal of visualized experiments : JoVE, e51138-e51138.

Xu, C., Fan, Z.P., Muller, P., Fogley, R., DiBiase, A., Trompouki, E., Unternaehrer, J., Xiong, F., Torregroza, I., Evans, T., Megason, S.G., Daley, G.Q., Schier, A.F., Young, R.A., Zon, L.I., 2012. Nanog-like regulates endoderm formation through the Mxtx2-Nodal pathway. Dev Cell 22, 625-638.

Xu, P., Zhu, G., Wang, Y., Sun, J., Liu, X., Chen, Y.G., Meng, A., 2014. Maternal Eomesodermin regulates zygotic nodal gene expression for mesendoderm induction in zebrafish embryos. J Mol Cell Biol 6.

Xu, Y., Yang, W., Wu, J., Shi, Y., 2002. Solution Structure of the First HMG Box Domain in Human Upstream Binding Factor. Biochemistry 41, 5415-5420.

Yamamoto, A., Amacher, S.L., Kim, S.H., Geissert, D., Kimmel, C.B., De Robertis, E.M., 1998. Zebrafish paraxial protocadherin is a downstream target of spadetail involved in morphogenesis of gastrula mesoderm. Development 125.

Yamashita, S., Miyagi, C., Fukada, T., Kagara, N., Che, Y.S., Hirano, T., 2004. Zinc transporter LIVI controls epithelial-mesenchymal transition in zebrafish gastrula organizer. Nature 429, 298-302.

Yan, Y.-L., Miller, C.T., Nissen, R., Singer, A., Liu, D., Kirn, A., Draper, B., Willoughby, J., Morcos, P.A., Amsterdam, A., Chung, B.-c., Westerfield, M., Haffter, P., Hopkins, N., Kimmel, C., Postlethwait, J.H., 2002. A zebrafish sox9 gene required for cartilage morphogenesis. Development 129, 5065.

Yang, B., Zhai, G., Gong, Y., Su, J., Han, D., Yin, Z., Xie, S., 2017. Depletion of insulin receptors leads to β-cell hyperplasia in zebrafish. Science Bulletin 62, 486-492.

Yang, C., Huang, M., DeBiasio, J., Pring, M., Joyce, M., Miki, H., Takenawa, T., Zigmond, S.H., 2000. Profilin enhances Cdc42-induced nucleation of actin polymerization. The Journal of cell biology 150, 1001-1012.

Yang, D., Lutter, D., Burtscher, I., Uetzmann, L., Theis, F.J., Lickert, H., 2014. miR-335 promotes mesendodermal lineage segregation and shapes a transcription factor gradient in the endoderm. Development 141, 514.

Yang, H., Zhou, Y., Gu, J., Xie, S., Xu, Y., Zhu, G., Wang, L., Huang, J., Ma, H., Yao, J., 2013. Deep mRNA sequencing analysis to capture the transcriptome landscape of zebrafish embryos and larvae. PloS one 8, e64058-e64058.

Yasuoka, Y., Suzuki, Y., Takahashi, S., Someya, H., Sudou, N., Haramoto, Y., Cho, K.W., Asashima, M., Sugano, S., Taira, M., 2014. Occupancy of tissue-specific cis-regulatory modules by Otx2 and TLE/Groucho for embryonic head specification. Nature communications 5, 4322-4322.

Ye, Z., Chen, Z., Sunkel, B., Frietze, S., Huang, T.H., Wang, Q., Jin, V.X., 2016. Genome-wide analysis reveals positional-nucleosome-oriented binding pattern of pioneer factor FOXA1. Nucleic Acids Res 44, 7540-7554.

Yiangou, L., Ross, A.D.B., Goh, K.J., Vallier, L., 2018. Human Pluripotent Stem Cell-Derived Endoderm for Modeling Development and Clinical Applications. Cell Stem Cell 22, 485-499.

Yuan, X., Song, M., Devine, P., Bruneau, B.G., Scott, I.C., Wilson, M.D., 2018. Heart enhancers with deeply conserved regulatory activity are established early in zebrafish development. Nat Commun 9, 4977.

Yun, A., Kang, B., Kim, C.Y., Jeong, D., Bae, D., Kim, E., Jung, H., Han, H., Jeon, H.-N., Kim, H., Cho, J.-W., Chung, M., Lee, M., Lee, S., Lee, S., Nam, S., Yang, S., Kim, Y., Lee, I., Kim, J.-H., 2017. TRRUST v2: an expanded reference database of human and mouse transcriptional regulatory interactions. Nucleic Acids Research 46, D380-D386.

Zaret, K.S., Carroll, J.S., 2011. Pioneer transcription factors: establishing competence for gene expression. Genes Dev 25, 2227-2241.

Zecchin, E., Conigliaro, A., Tiso, N., Argenton, F., Bortolussi, M., 2005. Expression analysis of jagged genes in zebrafish embryos. Developmental Dynamics 233, 638-645.

Zecchin, E., Filippi, A., Biemar, F., Tiso, N., Pauls, S., Ellertsdottir, E., Gnugge, L., Bortolussi, M., Driever, W., Argenton, F., 2007. Distinct delta and jagged genes control sequential segregation of pancreatic cell types from precursor pools in zebrafish. Dev Biol 301, 192-204.

Zeng, Q., Liu, S., Yao, J., Zhang, Y., Yuan, Z., Jiang, C., Chen, A., Fu, Q., Su, B., Dunham, R., Liu, Z., 2016. Transcriptome Display During Testicular Differentiation of Channel Catfish (Ictalurus punctatus) as Revealed by RNA-Seq Analysis. Biology of reproduction 95, 19.

Zhang, C., Basta, T., Klymkowsky, M.W., 2005. SOX7 and SOX18 are essential for cardiogenesis in Xenopus. Developmental dynamics : an official publication of the American Association of Anatomists 234, 878-891.

Zhang, D., Gates, K.P., Barske, L., Wang, G., Lancman, J.J., Zeng, X.-X.I., Groff, M., Wang, K., Parsons, M.J., Crump, J.G., Dong, P.D.S., 2017. Endoderm Jagged induces liver and pancreas duct lineage in zebrafish. Nature communications 8, 769-769.

Zhang, J., Chiodini, R., Badr, A., Zhang, G., 2011. The impact of next-generation sequencing on genomics. Journal of genetics and genomics = Yi chuan xue bao 38, 95-109.

Zhang, X., Guo, C., Chen, Y., Shulha, H.P., Schnetz, M.P., LaFramboise, T., Bartels, C.F., Markowitz, S., Weng, Z., Scacheri, P.C., Wang, Z., 2008. Epitope tagging of endogenous proteins for genome-wide ChIP-chip studies. Nature methods 5, 163-165.

Zhang, Z.H., Jhaveri, D.J., Marshall, V.M., Bauer, D.C., Edson, J., Narayanan, R.K., Robinson, G.J., Lundberg, A.E., Bartlett, P.F., Wray, N.R., Zhao, Q.-Y., 2014. A Comparative Study of Techniques for Differential Expression Analysis on RNA-Seq Data. PLOS ONE 9, e103207.

Zhao, S., Fung-Leung, W.-P., Bittner, A., Ngo, K., Liu, X., 2014a. Comparison of RNA-Seq and Microarray in Transcriptome Profiling of Activated T Cells. PLOS ONE 9, e78644.

Zhao, S., Zhang, Y., Gamini, R., Zhang, B., von Schack, D., 2018. Evaluation of two main RNA-seq approaches for gene quantification in clinical RNA sequencing: polyA+ selection versus rRNA depletion. Scientific Reports 8, 4781.

Zhao, S., Zhang, Y., Gordon, W., Quan, J., Xi, H., Du, S., von Schack, D., Zhang, B., 2015. Comparison of stranded and non-stranded RNA-seq transcriptome profiling and investigation of gene overlap. BMC genomics 16, 675-675.

Zhao, W., He, X., Hoadley, K.A., Parker, J.S., Hayes, D.N., Perou, C.M., 2014b. Comparison of RNA-Seq by poly (A) capture, ribosomal RNA depletion, and DNA microarray for expression profiling. BMC Genomics 15, 419.

Zhou, Q., Chipperfield, H., Melton, D.A., Wong, W.H., 2007. A gene regulatory network in mouse embryonic stem cells. Proceedings of the National Academy of Sciences 104, 16438.

Zhou, Q., Su, X., Jing, G., Chen, S., Ning, K., 2018. RNA-QC-chain: comprehensive and fast quality control for RNA-Seq data. BMC genomics 19, 144-144.

Zhou, Y., Williams, J., Smallwood, P.M., Nathans, J., 2015. Sox7, Sox17, and Sox18 Cooperatively Regulate Vascular Development in the Mouse Retina. PLoS ONE 10, e0143650.

Zhu, J., Fukushige, T., McGhee, J.D., Rothman, J.H., 1998. Reprogramming of early embryonic blastomeres into endodermal progenitors by a Caenorhabditis elegans GATA factor. Genes & development 12, 3809-3814.

Zorn, A.M., Wells, J.M., 2009. Vertebrate Endoderm Development and Organ Formation. Annual review of cell and developmental biology 25, 221-251.